FACULTAD DE CIENCIAS

DEPARTAMENTO DE QUÍMICA-FÍSICA APLICADA

TESIS DOCTORAL

Nutrigenómica: análisis experimental, integrativo y desarrollo de una plataforma de minería de datos

INSTITUTO MADRILEÑO DE ESTUDIOS AVANZADOS EN ALIMENTACIÓN

(IMDEA FOOD INSTITUTE)

Roberto Martín Hernández

Madrid 2019

FACULTAD DE CIENCAS

DEPARTAMENTO DE QUÍMICA-FÍSICA APLICADA

Memoria de Investigación presentada por ***Roberto Martín Hernández***

Para optar al grado de ***Doctor***

Por la ***Universidad Autónoma de Madrid***

Nutrigenómica: análisis experimental, integrativo y desarrollo de una plataforma de minería de datos

INSTITUTO MADRILEÑO DE ESTUDIOS AVANZADOS EN ALIMENTACIÓN
(IMDEA FOOD INSTITUTE)

Directores de la Tesis:

Dr. Alberto Dávalos
Investigador Senior del Instituto IMDEA Alimentación
Prof. Guillermo Reglero
Catedrático de Ciencias de la Alimentación de la UAM

D. **ALBERTO DÁVALOS HERRERA**, DOCTOR EN FARMACIA POR LA UNIVERSIDAD COMPLUTENSE DE MADRID E INVESTIGADOR SENIOR DEL INSTITUTO IMDEA ALIMENTACIÓN, Y D. **GUILLERMO REGLERO RADA**, DOCTOR EN CIENCIAS QUÍMICAS POR LA UNIVERSIDAD AUTÓNOMA DE MADRID Y CATEDRÁTICO DE LA UNIVERSIDAD AUTÓNOMA DE MADRID

INFORMAN:

Que el presente trabajo titulado: "Nutrigenómica: análisis experimental, integrativo y desarrollo de una plataforma de minería de datos" presentada por Don Roberto Martín Hernández, para optar al grado de Doctor en Ciencias de la Alimentación por la Universidad Autónoma de Madrid, ha sido realizado en el Instituto Madrileño de Estudios Avanzados en Alimentación (IMDEA Alimentación) bajo nuestra dirección, y cumple las condiciones requeridas por la Universidad Autónoma de Madrid para optar al Título de Doctor.

Y para que así conste, firman el presente informe:

Dr. Alberto Dávalos

Prof. Guillermo Reglero

El autor agradece la financiación de la Agencia Estatal de Investigación y Fondos Europeos FEDER (AGL2016-78922-R) a través del proyecto titulado: “Modulación terapéutica de ARNs no codificantes a través de componentes bioactivos de la dieta: impacto sobre la regulación fisiopatológica del metabolismo lipídico intestinal”. También la de la Fundación Ramón Areces (CIVP18A3888) a través del proyecto titulado: “Modulación de exosomas transportadores de miRNAs y lncRNAs para la comunicación intercelular como herramienta terapéutica frente a la dislipidemia”.

El autor agradece también a IMDEA Alimentación el contrato financiado por fondos FEDER Europeos así como fondos de la Comunidad de Madrid.

"Our lives are defined by opportunities,

even the ones we miss"

Eric Roth, The Curious Case of Benjamin Button

A mis seres queridos

AGRADECIMIENTOS

Cuando llegué al instituto IMDEA Alimentación, proveniente de una empresa de Bioinformática, mis conocimientos sobre los efectos de la nutrición en la salud eran prácticamente nulos. Durante estos años he aprendido mucho sobre las ciencias de la alimentación, y hoy, cuando miro atrás, me siento abrumado por haber tenido la posibilidad de trabajar con científicos de tan alto nivel, expertos en la materia, y a quienes me gustaría mostrar mi agradecimiento a continuación-

Los dos primeros científicos a los que quiero mostrar mi agradecimiento son a mis dos directores de tesis, el Dr. Alberto Dávalos y el profesor Guillermo Reglero. Sin duda Alberto es el artífice de la realización de esta tesis doctoral. Fue él quien me sugirió aplicar la herramienta bioinformática a las ciencias de la alimentación, y juntos concebimos el proyecto que hoy se ve traducido en tesis doctoral. Gracias por tu constante apoyo, la motivación que me has aportado, y por contagiarme de tu rigor e incansable curiosidad científica. Lo considero sin duda mi mentor en el mundo académico. En cuanto a Guillermo, como director del centro, admiro profundamente su infatigable trabajo, desempeño y dedicación para el continuo desarrollo del instituto IMDEA Alimentación.

En segundo lugar, son muy importantes para mí los 3 compañeros con los que hasta ahora he tenido el honor de compartir el despacho de Bioinformática y Bioestadística, y que tanto me han enseñado. Por orden cronológico, estas 3 personas son; 1) Don Jesús Herranz, maestro de la estadística y programación en R de la vieja escuela. Ha sido muy agradable compartir nuestra debilidad por la cultura musical africana y jamaiquina. También conocido como DJ Tubabu, es el fundador y director del programa de radio “AfricaPachanga”. 2) Don Manuel Calas, experto técnico informático y compañero de desayunos. Gracias por contagiarme con tu inquebrantable sonrisa y buen humor. 3) El Dr. Gonzalo Colmenarejo, maestro de la modelización de cualquier tipo de dato y programación en distintos lenguajes. Gracias a los 3 por haber hecho tan agradables los momentos que hemos, y seguimos, compartiendo. De todos ellos he procurado aprender de su dilatada experiencia laboral y profundos conocimientos técnicos.

En tercer lugar me gustaría mostrar mi agradecimiento a las integrantes de la plataforma GENYAL: Isabel, Viviana, Elena Aguilar, Elena Marcos y Rocío. Expertas nutricionistas,

trabajadoras incansables, con su continuo buen humor y disponibilidad para ayudar en cualquier cosa. También mostrar mi agradecimiento a Susana Molina, responsable del laboratorio de GENYAL.

En cuarto lugar, quiero agradecer a las integrantes de los grupos de investigación de "Bioactive Food Ingredients" y "Epigenetics of Lipid Metabolism" por todos los buenos momentos que hemos compartido juntos, tanto dentro como fuera del instituto. Estas son Judit, Carmen, Mayka, Diana, Lorena y por supuesto João.

En quinto lugar, no puedo dejar de mencionar a los investigadores más experimentados del instituto, cada uno experto en su materia, y de los que tanto he aprendido teniendo el honor de analizar sus datos. Estos son el Prof. José María Ordovás, el Dr. Francesco Visioli, el Dr. Pablo José Fernández, por supuesto mi director de tesis el Dr. Alberto Dávalos, el Prof. Alfredo Martínez y la Dra. Ana Ramírez de Molina.

En sexto lugar, mis agradecimientos van para todos mis compañeros imdeanos, tanto técnicos como estudiantes pre doctorales y post doctorales, siempre amables y disponibles para ayudar en cualquier momento de manera desinteresada. Estos son Lidia, Victor, Rodrigo (maestro del paquete Office), Ruth Sánchez, Marta Gómez, Marga, Teo, Lara, Josune, Jorge, Silvia, Lamia, Mónica, Luis Filipe y Marta Barradas. También agradecer su amabilidad y disponibilidad al personal de gestión, secretaría y comunicación del centro: Inmaculada Galindo, Gema, Jowita y Sara. Lo mismo para el personal de la recepción del centro, Sara y Esther. Sin olvidar al personal de limpieza y mantenimiento del centro, especialmente Julia y Rubén, auténticos guardianes del orden imdeano en ambas alas Este y Oeste. Igualmente gracias a Eduardo, técnicamente el mejor programador que conozco, por todas sus enseñanzas concisas sobre RoR, Javascript y otras muchas.

Finalmente esta tesis va dedicada a mis padres, Maite y Roberto, a mi tía y madrina María Luisa, a mis tías Maite, Pilar y Raquel, así como a todos mis primos hermanos Pablo, Mónica, Minerva, David, Jordi e Ignacio, y a mis primitas Nora y Martina.

ÍNDICE

CONTENIDO

ABREVIATURAS

ADN: deoxyribonucleic acid / ácido desoxirribonucleico

ADNc: complementary deoxyribonucleic acid / ácido desoxirribonucleico complementario

Akt: protein kinase B / proteína quinasa B

AMF: amorfrutin / amorfrutina

ARN: ribonucleic acid / ácido ribonucleico

ARNm: messenger ribonucleic acid / ácido ribonucleico mensajero

CA: carnosic acid / ácido carnósico

CMap: connectivity map / mapa de conectividad

ECV: enfermedad cardiovascular

EGCG: epigallocatechin gallate / epigalocatequina-galato

EGF: epidermal growth factor / factor de crecimiento epidérmico

ES: enrichment score / puntuación de enriquecimiento

FDR: false discovery rate / tasa de descubrimiento falso

HT: hydroxytyrosol / hidroxitirosol

I3C: indole-3-carbinol / indol-3-carbinol

GEO: gene expression omnibus / ómnibus de expresión génica

GSEA: gene set enrichment analysis / análisis de enriquecimiento de conjuntos de genes

LB: Lactobacillus

Log2 FC: base 2 logarithm fold change / cambio proporcional en logaritmo en base 2

miARN: micro-ribonucleic acid / micro ácido ribonucleico

MUFA: monounsaturated fatty acid / ácido graso monoinsaturado

NES: normalized enrichment score / puntuación de enriquecimiento normalizada

NGS: next generation sequencing / secuenciación de nueva generación

NF-κB: nuclear factor kappa-light-chain-enhancer of activated B cells / factor nuclear potenciador de las cadenas ligeras kappa de las células B activadas

PCR: polymerase chain reaction / reacción en cadena de la polimerasa

PDGF: platelet-derived growth factor / factor de crecimiento derivado de plaquetas

PI3K: phosphoinositide 3-kinase/ fosfoinositida-3-quinasa

PPAR: peroxisome proliferator-activated receptor / receptores activados por proliferadores de peroxisomas

PUFA: polyunsaturated fatty acid / ácido graso poliinsaturado

RVT: resveratrol

RMA: robust multi-array average / media robusta multi-array

RSM: rosemary / romero

Sirt1: sirtuin 1 / sirtuina 1

SFP: sulforaphane / sulforafano

TCT: tocotrienol

TNFα: tumor necrosis factor alpha / factor de necrosis tumoral alfa

TRVT: trans-resveratrol

UTR: untranslated region / región no traducida

VEGF: vascular endothelial growth factor / factor de crecimiento endotelial vascular

WFNA: witaferin A / witaferina A

RESUMEN/SUMMARY

La Nutrigenómica, ciencia que estudia el potencial de los alimentos y sus compuestos bioactivos para alterar la expresión de los genes, podría explicar la capacidad de la dieta para modular nuestra salud. Plataformas especializadas como Gene Expression Omnibus (GEO) reúnen multitud de datos de expresión génica generados por tecnologías ómicas, y contienen resultados de experimentos que analizan los cambios de expresión génica en células humanas tras su tratamiento con distintos alimentos y compuestos bioactivos. El análisis integrativo de estos datos ofrece la posibilidad de profundizar en el conocimiento de las bases moleculares que gobiernan el binomio dieta-salud. Esta tesis trata de contribuir al estudio del potencial nutrigenómico de los alimentos y sus compuestos bioactivos. Inicialmente, se ha demostrado como el consumo de una dieta suplementada en hidroxitirosol, principal fitoquímico fenólico bioactivo del aceite de oliva virgen, es capaz de regular la expresión de cuatro micro ARN's *in vivo* en el hígado de ratón, con potenciales implicaciones biológicas. Tras ello, se han recopilado y analizado datos resultantes de experimentos de nutrigenómica, disponibles en GEO, para construir una base de datos de expresión diferencial de genes. El análisis integrativo de esta base de datos ha permitido identificar una firma molecular de 18 genes con potenciales propiedades anticancerígenas. Finalmente, se ha completado la base de datos inicial con nuevos datos y definido las firmas moleculares de los experimentos incluidos, para desarrollar una plataforma de minería de datos en nutrigenómica. La plataforma se presenta en forma de una aplicación web, accesible públicamente (www.nutrigenomedb.org). Mediante el uso de tablas y gráficos interactivos, esta plataforma permite explorar el nivel de expresión diferencial de genes a nivel celular en respuesta al tratamiento con alimentos y sus compuestos bioactivos. Mediante un algoritmo de comparación de patrones de expresión, las firmas moleculares externas pueden compararse con las incluidas en la base de datos de nutrigenómica, lo que permite identificar mecanismos moleculares en común que explicarían los efectos beneficiosos de determinados alimentos. A través de un caso de uso, se demuestra como la aplicación desarrollada permite conectar una firma molecular provocada por el Amlodipino, un fármaco usado para el tratamiento de la hipertensión, con una firma molecular obtenida tras un tratamiento con un extracto de romero. Además el análisis funcional de los genes implicados en esta conexión identifica la represión de genes involucrados en actividades de transporte transmembrana de iones como el principal mecanismo molecular responsable de la conexión identificada.

Nutrigenomics, the science in charge of studying the effects of foods and their bioactive compounds on gene expression, offers great possibilities for explaining how diet is able to modulate human health. Multiple omic experiments, available at Gene Expression Omnibus (GEO) database, have generated gene expression data following treatment of human cell lines with different foods and their bioactive compounds. Exploration of such data in an integrative manner offers excellent possibilities for gaining insights into the molecular basis governing the diet-health binomial. This doctoral thesis focuses on the nutrigenomic potential of foods and their bioactive compounds. We started showing that hydroxytyrosol, the main bioactive compound from virgin olive oil, is able to modulate the expression of 4 micro RNA's *in vivo* on mouse liver, with potential biological outcomes. Then we have collected and analyzed nutrigenomics experiments publicly available in GEO database, in order to build a gene expression database. Integrative analysis of such database allowed us to identify a molecular signature of 18 genes with potential anticancer properties. Finally, this database has been updated with new data, and molecular signatures defining the gathered experiments have been obtained, in order to build up a nutrigenomics data mining platform. Such a platform is presented as an open web application (www.nutrigenomedb.org). Through its web interface, users are able to explore differential gene expression data at cellular level, in response to different treatments with a variety of foods and their bioactive compounds, by using data tables and interactive visualizations. In addition, external gene signatures can be connected with hosted nutrigenomics molecular signatures using a gene pattern-matching algorithm. We further demonstrate how the application is able to connect a molecular signature triggered by a cellular treatment with Amlodipine, a drug used to treat hypertension, with a molecular signature corresponding to a cellular treatment with a rosemary extract. A functional analysis of the genes connecting both treatments identifies the downregulation of a set of genes related to ion transmembrane transport activities as the main underlying molecular mechanism in common.

INTRODUCCIÓN

Entre todos los factores ambientales que afectan a la homeostasis de los organismos, la dieta es uno de los más importantes. Se puede decir que nos encontramos inmersos en la era de la nutrición molecular, dado que tenemos a nuestra disposición las herramientas para estudiar y caracterizar las interacciones entre nuestros genes y los nutrientes que obtenemos de los alimentos. La posibilidad de tratar a los ingredientes de los alimentos como un conjunto de moléculas, las cuales utilizadas en dosis específicas, puedan tener propiedades beneficiosas para la salud, sería un paso fundamental para promover la prevención de las enfermedades crónicas más comunes a través de la dieta.

Existen principalmente dos ciencias que se ocupan de este campo de estudio. Por un lado se encuentra la nutrigenética, ciencia que estudia la influencia de las variaciones genéticas en la respuesta del organismo a los nutrientes, y por otro lado está la nutrigenómica, que abarca el estudio de la influencia de los nutrientes en la expresión de los genes. Este trabajo de investigación se centra en la nutrigenómica. La primera parte tratará del análisis de datos de nutrigenómica utilizando las herramientas moleculares más actuales. En la segunda parte se analizarán datos de nutrigenómica de manera integrativa, mediante diferentes técnicas de minería de datos, con el fin de extraer nuevos conocimientos sobre los mecanismos moleculares básicos que provoca el tratamiento de células humanas con distintos nutrientes y compuestos alimenticios. Finalmente se presentará el desarrollo de una plataforma web, públicamente disponible en línea, para la exploración y minería de datos nutrigenómicos humanos, con el fin de indagar en las propiedades saludables observadas que han demostrado determinados alimentos y compuestos bioactivos.

1. BINOMIO DIETA-SALUD

La dieta representa un factor ambiental que influye de manera directa sobre la salud. Es un concepto muy antiguo cuyas primeras referencias apuntan a la Grecia antigua de Hipócrates. Enfermedades de gran incidencia social, tales como la obesidad, la enfermedad cardiovascular (ECV), la diabetes o el cáncer, se deben a interacciones complejas entre diversos genes y a factores ambientales como la dieta. Prevenir la aparición de estas enfermedades, o bien mejorar la salud, a través de la dieta, representa un importante reto para la comunidad científica del siglo XXI.

1.1. ESTUDIOS EPIDEMIOLÓGICOS OBSERVACIONALES

Los estudios epidemiológicos observacionales en nutrición han sido hasta el momento la herramienta más importante en el estudio de la influencia de la dieta sobre la salud, aportando las primeras evidencias de esta relación. En estos estudios se constituyen cohortes de individuos sanos a los que, tras definir su exposición inicial a distintos factores, se les realiza un seguimiento durante

un largo periodo de tiempo para evaluar su evolución (1). Los seguimientos se realizan principalmente mediante cuestionarios de frecuencia de consumo de alimentos y anotaciones de las rutinas alimentarias por parte de los voluntarios. Generalmente se trata de estudios retrospectivos, ya que se trabaja con 2 bloques distintos de información, uno obtenido al inicio del estudio y otro generado al cabo de un periodo de tiempo determinado, pudiendo extraer así distintas informaciones prolongando el segundo bloque en distintas longitudes de tiempo. Es importante estudiar grandes cohortes de individuos, ya que de ello depende la robustez estadística de los resultados obtenidos. La presencia de potenciales factores de confusión, variables presentes no estudiadas que pueden explicar la asociación entre la variable estudiada y el efecto observado, representa la principal desventaja de los estudios epidemiológicos.

Uno de los estudios epidemiológicos observacionales que más ha contribuido a evidenciar la relación entre salud y dieta es el Estudio Framingham del Corazón, iniciado en el año 1948 con una cohorte de 5.209 hombres y mujeres**,** y que continúa en la actualidad (2). Fue uno de los primeros estudios epidemiológicos centrado en la ECV. Antes del inicio de este estudio, la comunidad científica asumía que parámetros fisiológicos tales como una presión arterial elevada, altos niveles de colesterol en sangre y los procesos de aterosclerosis, eran consecuencia inevitable del envejecimiento y no se podían prevenir. Sin embargo los primeros resultados del estudio Framingham demostraron que los altos niveles de colesterol en sangre representan un importante factor de riesgo en vista a padecer ECV, así como que estos factores de riesgo se podían reducir a través de la dieta. Los resultados del Estudio Framingham del Corazón han sido confirmados posteriormente en distintos trabajos (3,4), demostrando la clara influencia que presenta la dieta ante el riesgo de padecer ECV. De hecho, una de las contribuciones más interesantes del estudio Framingham ha sido el desarrollo de un algoritmo para el cálculo de una puntuación que permite predecir, en función del sexo, la probabilidad de padecer ECV a diez años vista (**Figura 1**).

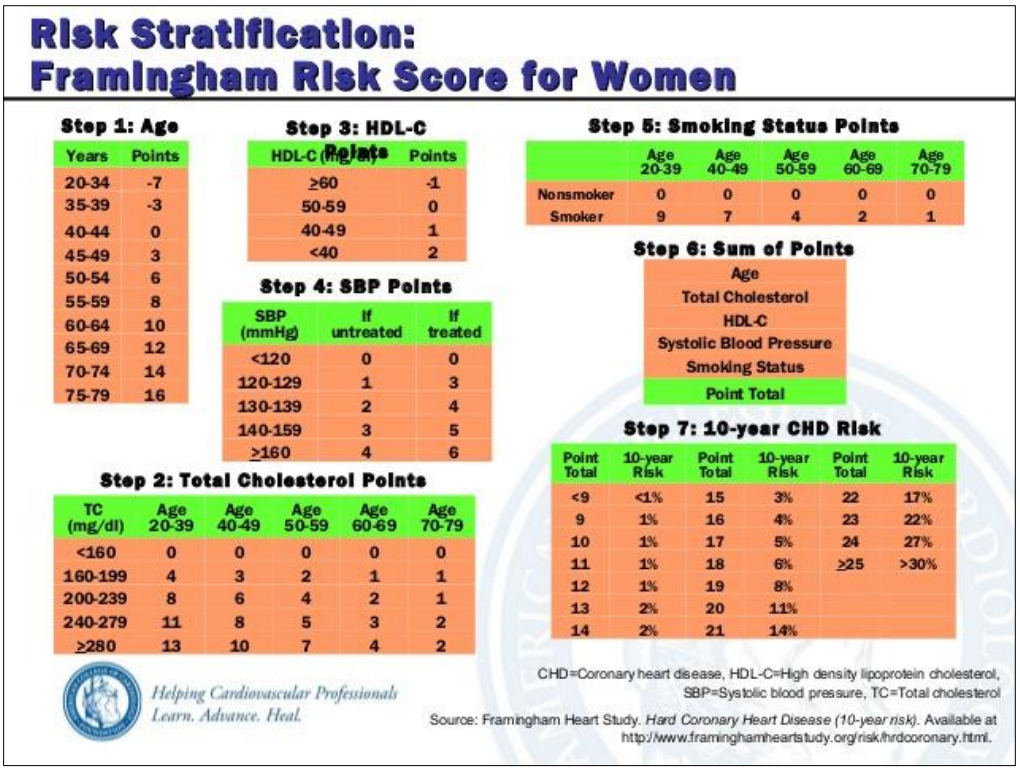


Figura 1: Ilustración del sistema de puntuación obtenido en el estudio Framingham, utilizado para calcular la probabilidad (expresada en porcentaje) de sufrir enfermedad cardiovascular en un periodo de 10 años en el caso de las mujeres. Imagen obtenida de la página web principal del estudio Framingham.

Desde entonces, la investigación en epidemiología nutricional ha continuado obteniendo resultados a partir de amplios estudios epidemiológicos observacionales, realizados en distintas cohortes. Se han establecido asociaciones de cierta robustez entre la frecuencia de consumo de alimentos específicos con la incidencia de enfermedades cardiovasculares, metabólicas e incluso algunos tipos de cáncer. En la **tabla 1** se encuentran recopilados una lista de importantes resultados sobre la relación entre dieta y salud, obtenidos en epidemiología nutricional a partir de estudios observacionales de cohortes, y publicados en dos de las revistas científicas más prestigiosas del área de medicina ("Journal of the American Medical Association" o JAMA, y "The New England Journal of Medicine" o NEJM).

Los estudios presentados en la **tabla 1** han establecido asociaciones entre la dieta ingerida y la incidencia de distintas enfermedades crónicas. Sin embargo a día de hoy es difícil aportar una explicación del efecto observado a nivel molecular, así como distinguir cuáles son las moléculas bioactivas de los alimentos que otorgan al organismo un efecto protector frente a enfermedades crónicas. Como ejemplo de la complejidad de un alimento podemos utilizar el aceite de oliva, una sustancia grasa de origen vegetal que contiene una amplia variedad de compuestos en distintas concentraciones: ácidos grasos, triglicéridos, esteroles, tocoferoles y polifenoles. Todos estos compuestos pueden ejercer diversas funciones moleculares en destinos celulares diferentes.

Nombre de estudio o cohorte	Factor dietario	Tamaño muestra	País	Asociación significativa	Pubmed ID
Predimed	Dieta mediterránea	7.447 h/m	España	Reducción del riesgo de muerte por enfermedad cardiovascular	23432189
N.D.	Consumo de frutas	512.891 h/m	China	Reducción de presión arterial, niveles de azúcar en sangre y menor riesgo de muerte por enfermedad cardiovascular	27050205
Nurses' Health Study+Health Professionals Follow-Up Study	Consumo de nueces	76.464 + 42.498 h/m	EE.UU.	Reducción de mortalidad total	24256379
National Institutes of Health-AARP Diet and Health Study	Consumo de café	229.119 + 173.141 h/m	EE.UU.	Reducción de mortalidad total	22591295
N.D.	Dieta mediterránea	22.043 h/m	Grecia	Reducción de mortalidad total	12826634
Chicago Western Electric Study	Consumo de pescado	1.822 hombres	EE.UU.	Reducción de muerte por enfermedad coronaria	9091800
Health Professionals Follow-Up Study	Vitamina E	39.910 hombres	EE.UU.	Reducción del riesgo de mortalidad por enfermedad coronaria	8479464
Predimed	Vitamina K1	5.860 h/m	España	Reducción del riesgo de sufrir cataratas en población anciana	28494067
N.D.	Frutas, verduras	8.104 + 8.516 h/m	EE.UU.	Disminución de la proporción de muertes por infarto, accidente cerebro-vascular y diabetes tipo 2	5852674
Amyotrophic Lateral Sclerosis Multicenter	Antioxidantes, carotenos, frutas y verduras	302 pacientes	EE.UU.	Mejora de funcionalidad en pacientes con Esclerosis lateral amiotrófica	27775751
Nurses' Health Study+Health Professionals Follow-Up Study	Proteína vegetal	131.342 h/m	EE.UU.	Reducción de mortalidad en general	27479196
Nurses' Health Study+Health Professionals Follow-Up Study	Ácidos omega-3 marinos	173.229 h/m	EE.UU.	Reducción de incidencia de cáncer colorrectal	27148825
Adventist Health Study 2	Dieta vegetariana	96.354 h/m	EE.UU.	Reducción de incidencia de cáncer colorrectal	25751512
Nurses' Health Study+Health Professionals Follow-Up Study	Alimentos integrales	74.341 m. + 43.744 h.	EE.UU.	Reducción de mortalidad por enfermedad cardiovascular	25559238
Shanghai Breast Cancer Survival Study	Consumo de soja	5.042 mujeres	China	Reducción de riesgo de muerte y recurrencia en cáncer de mama	19996398
N.D.	Dieta occidental	1.009 pacientes	EE.UU.	Mayor recurrencia y mortalidad entre pacientes	17699009
Swedish Mammography Cohort	Pescado azul	61.433 mujeres	Suecia	Reducción de la incidencia de adenocarcinoma renal	16985229
N.D.	Fitoestrógenos	1.674 pacientes	EE.UU.	Reducción del riesgo de cáncer de pulmón	16189362
Rotterdam study	Betacaroteno, vitaminas C y E, zinc	5836 h/m	Holanda	Reducción del riesgo de degeneración macular asociada a la edad	16380590
Nurses' Health Study	Mantequilla de nueces y cacahuete	83.818 mujeres	EE.UU.	Reducción del riesgo de padecer diabetes tipo 2	12444862
Cardiovascular Health Study	Fibra de cereal	3.588 h/m	EE.UU.	Reducción de la incidencia de enfermedad cardiovascular	12672734
Health Professional Follow-up Study	Consumo de pescado	43.671 hombres	EE.UU.	Reducción del riesgo de accidente cerebrovascular	12495393
Nurses' Health Study	Alimentos integrales	75.521 mujeres	EE.UU.	Reducción del riesgo de accidente cerebrovascular	11000647
Honolulu Heart Program	Café y cafeina	8.004 hombres	Japón-EE.UU.	Menor incidencia de enfermedad de Parkinson	10819950

Tabla 1: Listado de estudios epidemiológicos que han encontrado asociaciones estadísticamente significativas entre salud y dieta. Los resultados han sido publicados en las revistas del área de medicina JAMA y NEJM.

1.2 ENSAYOS CLÍNICOS: EL EJEMPLO DE LA DIETA MEDITERRÁNEA

En epidemiología nutricional, los resultados observados en estudios observacionales de cohortes suelen ser investigados posteriormente en ensayos clínicos. A diferencia de los estudios de cohorte, los ensayos clínicos son estudios experimentales en los cuales el investigador diseña el estudio utilizando grupos aleatorizados de individuos de tratamiento y control. Los ensayos clínicos aleatorizados permiten eliminar muchos de los sesgos que presenta el estudio de cohortes, dado el control absoluto que el investigador ejerce respecto a la intervención nutricional realizada. La posibilidad de reclutar individuos mediante determinados criterios de inclusión permite homogeneizar al máximo los grupos estudiados. De esta manera se consiguen minimizar los potenciales factores de confusión que presentan los estudios observacionales, y las asociaciones encontradas pueden atribuirse directamente al efecto de la intervención nutricional. Sin embargo presentan la desventaja de ser estudios bastante costosos económicamente, y a diferencia de los estudios observacionales de cohorte, los estudios clínicos sólo pueden abordar un reducido número de hipótesis. De la misma manera, los resultados obtenidos dependen en gran medida del compromiso del voluntario mediante su adherencia a la intervención, así como de la honestidad de los voluntarios a la hora de la recogida de datos mediante cuestionarios. Dado que los estudios de cohortes han demostrado los amplios beneficios para la salud que presenta la exposición a elementos propios de la dieta mediterránea, ésta última es una de las dietas más estudiadas mediante ensayos clínicos nutricionales.

El concepto de dieta mediterránea, desarrollado a principios del año 1960, hace referencia a un estilo de vida equilibrado. Este estilo de vida engloba las dietas y formas de cocinar típicas de zonas bañadas por el mar mediterráneo tales como la isla de Creta, Grecia o el sur de Italia, así como las costumbres de dormir la siesta y comer en grupo. Estudios observacionales han demostrado que estos lugares del mediterráneo presentan una de las mayores esperanzas de vida a nivel mundial, con bajos niveles de muertes por accidente cardiovascular en relación a otros lugares del mundo , así como una menor mortalidad por cáncer y enfermedades neurodegenerativas (5). La dieta mediterránea se caracteriza por su alto contenido en alimentos integrales, frutas y verduras, pescados, frutos secos y aceite de oliva. Esta composición se traduce en aportes moleculares específicos tales como fitoquímicos, antioxidantes, ácidos grasos esenciales y fibra alimentaria. A continuación se detallan los resultados obtenidos en importantes estudios observacionales y ensayos clínicos con elementos de la dieta mediterránea sobre determinadas enfermedades. Nótese que aunque la dieta mediterránea es un concepto amplio (ver arriba) los estudios siguientes solo consideran en su mayoría la parte nutricional (dieta *per se*) del concepto de dieta mediterránea.

1.2.1 Dieta mediterránea y enfermedad cardiovascular

La ECV es el término utilizado para problemas relacionados con el corazón y los vasos sanguíneos. Estos problemas a menudo se deben a la acumulación de sustancias grasas y colesterol en el interior de las arterias, o aterosclerosis. Cuando las arterias resultan obstruidas, aumentan las posibilidades de sufrir un ataque cardíaco o un accidente cerebrovascular. La ECV incluye muchos otros trastornos tales como la hipertensión arterial, aneurismas, cardiopatías, fibrilación auricular o insuficiencia cardiaca. La ECV representa la principal causa de mortalidad en los países desarrollados. Como se ha explicado anteriormente, la dieta representa un importante factor de riesgo en vista a sufrir ECV.

El estudio más importante sobre la dieta mediterránea se ha realizado en España, siguiendo durante casi 5 años a una cohorte de 7.447 personas con alto riesgo de padecer ECV. En un ensayo clínico, se evaluaron 2 dietas de tipo Mediterráneo; la primera aumentando el consumo de aceite de oliva virgen extra, y la segunda aumentando el consumo de nueces y frutos secos. Tras finalizar el estudio y analizar los datos obtenidos, se concluyó que ambas dietas disminuían significativamente la incidencia de problemas cardiovasculares en la muestra estudiada en comparación con el grupo control, el cual consumió una dieta baja en grasa. El artículo original de 2013 fue retractado, debido a inconsistencias en la aleatorización de algunos voluntarios. Sin embargo se volvieron a publicar los nuevos resultados del estudio en 2018 tras realizar las correcciones pertinentes, y éstos seguían otorgando un efecto protector de la dieta Mediterránea a nivel cardiovascular (6).

Otro ensayo clínico realizado en Francia, bautizado como Medi-RIVAGE, examinó el efecto de una dieta mediterránea rica en cereales, frutos secos, pan integral, frutas, verduras y pescado, siendo el aceite de oliva la principal fuente de grasa. Esta dieta fue consumida durante 3 meses por 212 voluntarios con un riesgo moderado de sufrir ECV. Se observó una reducción significativa de colesterol, triglicéridos e insulina en sangre, siendo esto favorable para la prevención de ECV. Es importante señalar que en el mismo estudio también se evaluó el efecto del consumo de otra dieta baja en grasas, caracterizada por un consumo restringido de carnes animales, abundante pescado, frutas, verduras y aceites vegetales. Tras el consumo de la dieta baja en grasas, también se lograron reducir los factores de riesgo característicos de la ECV en la misma medida que tras el consumo de la dieta mediterránea (7).

1.2.2 Dieta mediterránea y síndrome metabólico

El síndrome metabólico se define como un conjunto de condiciones que aumentan las posibilidades de que un individuo desarrolle ECV y diabetes de tipo 2. Para su diagnóstico, un individuo debe presentar al menos tres de las siguientes condiciones:

- Hipertensión arterial
- Altos niveles de glucosa en sangre
- Altos niveles de triglicéridos en sangre
- Bajos niveles de colesterol HDL en sangre
- Exceso de grasa abdominal

En un estudio realizado a principios del año 2000 en Nápoles (Italia), con una muestra de 180 pacientes con síndrome metabólico y de 2 años de duración, se observó que el grupo de pacientes adherido a la dieta mediterránea presentaba una disminución significativa de marcadores de inflamación en el suero sanguíneo. Concretamente se observaron menores niveles de la proteína C reactiva de alta sensibilidad (hsCRP) e interleucinas 6 (IL-6), 7 (IL-7), 18 (IL-18), además de una menor resistencia a la insulina. Este estudio llegó a la conclusión que la dieta mediterránea reduce la prevalencia del síndrome metabólico y por consecuencia el riesgo de padecer ECV (8). Otro importante ensayo clínico en 459 pacientes con hipertensión arterial, realizado en Estados Unidos, demostró que una dieta rica en frutas, verduras y baja en grasa, permite disminuir de manera significativa la presión arterial (9).

A nivel observacional, el efecto de la dieta sobre la incidencia del síndrome metabólico también fue evaluado sobre una muestra de los descendientes del estudio de Framingham citado anteriormente ("Framingham Offspring Study"). En una muestra 2.834 descendientes, se estudió la relación entre dietas a base de carbohidratos y fibras con la resistencia a la insulina y el síndrome metabólico. Así se obtuvo la conclusión de que el incremento del consumo de alimentos integrales reduce el riesgo de desarrollar síndrome metabólico (10).

1.2.3 Dieta mediterránea y cáncer

El cáncer es un conjunto de enfermedades que se caracteriza por una división descontrolada de las células y que además tienen la capacidad de propagarse hacia otras zonas del cuerpo, proceso denominado metástasis. Representa la segunda causa de muerte en el mundo tras la ECV. Se estima que entre el 5-10% de los cánceres están relacionados con la genética del individuo, mientras que los factores ambientales explicarían entre el 90-95% de los casos restantes (11).

Distintos estudios observacionales han atribuido propiedades protectoras contra el cáncer a la dieta mediterránea (5). El estudio EPIC ("European Prospective Investigation Into Cancer and

Nutrition") consiguió establecer, entre los años 1992 y 2000, una de las cohortes más grandes del mundo, incluyendo hasta 520.000 individuos de entre 25 y 75 años de edad, a través de varios centros repartidos por toda Europa, con el objetivo de evaluar la relación entre nutrición y cáncer. Se utilizaron cuestionarios al comienzo del estudio, para establecer la adherencia de los individuos de la cohorte a la dieta mediterránea, evaluando la ingesta de alimentos tales como frutas, frutos secos, legumbres, cereales, pescados, carnes e incluso alcohol. El análisis de los cuestionarios permitió cuantificar la adherencia de los individuos a la dieta mediterránea a través de una puntuación. Al finalizar el periodo de seguimiento, alrededor del año 2005, y tras cuantificar los eventos de cáncer surgidos en la cohorte, se realizaron análisis de supervivencia a partir de los datos obtenidos. Los resultados demostraron una fuerte asociación entre el consumo de una dieta mediterránea y un menor riesgo de incidencia de cáncer. Esta asociación fue todavía más robusta entre los individuos fumadores (12). A nivel del estudio de recaída en la enfermedad, estudios observacionales en pacientes han confirmado que un mayor consumo de una dieta abundante en carnes rojas, carnes procesadas y dulces está asociado a mayores tasas de recaída (13).

Sin embargo el estudio de la relación entre dieta y cáncer resulta difícil de descifrar mediante el uso ensayos clínicos. El hecho de incluir a personas con una enfermedad tan grave en un ensayo clínico no es una práctica médica prudente. Si bien los pacientes de cáncer suelen mostrar una gran disposición para probar terapias complementarias basadas en cambios del estilo de vida, tales como cambios en la dieta, es difícil que cumplan los criterios de inclusión en este tipo de ensayos, así como que cuenten con el beneplácito médico (14). Dada la dificultad de relacionar la aparición de cáncer con la dieta mediante estudios clínicos, sí es posible estudiar el efecto de la dieta sobre la recaída en la enfermedad de pacientes sanos tras una terapia inicial, aunque los datos obtenidos a partir de estos estudios todavía no son abundantes (14,15).

1.3 Discrepancias entre estudios observacionales y clínicos

En las últimas décadas se han generado gran cantidad de resultados mediante ambos tipos de estudios, y muchas veces han resultado ser contradictorios. Sin embargo es importante considerar las limitaciones en los diseños de los estudios observacionales y clínicos. A diferencia de los estudios clínicos, los estudios observacionales padecen de la fuerte influencia de factores de confusión. Pero hay que tener en cuenta que, debido a limitaciones técnicas y económicas, los estudios clínicos se realizan con individuos de poblaciones específicas, con poca diversidad étnica, y su duración es más limitada en el tiempo que la de los estudios observacionales. Con lo cual la falta de asociación en un estudio clínico de nutrición no debería invalidar la asociación observada en un

amplio estudio observacional, sino más bien contribuir a mejorar los diseños experimentales de los estudios clínicos (16).

2. MECANISMOS MOLECULARES DE NUTRIGENÓMICA

Los alimentos ingeridos a través de la dieta proporcionan al organismo una gran cantidad de sustancias biológicamente activas. Estas pueden ejercer una amplia variedad de funciones en distintos destinos celulares, con potenciales beneficios para la salud. Los componentes de la dieta pueden alterar la expresión de los genes directa o indirectamente (**Figura 2**). A nivel celular, los nutrientes principalmente pueden:

• Actuar directamente como ligandos para activar factores de transcripción, iniciando rutas específicas de señalización celular.

• Influir en las rutas de señalización celular, modificando la expresión de los genes.

• Incorporarse en distintas rutas metabólicas primarias o secundarias, alterando la concentración de sustratos o intermediarios.

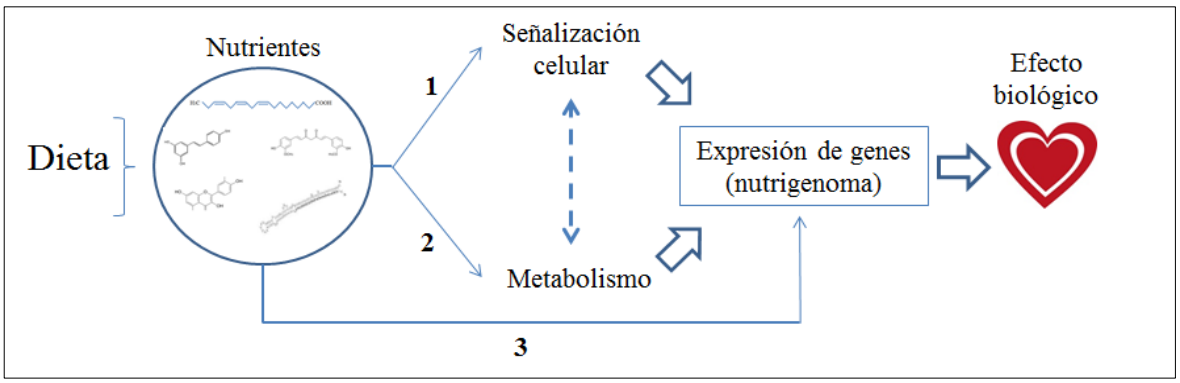


Figura 2: Esquema de las tres vías por las que pueden actuar los componentes de los alimentos para alterar el nutrigenoma. 1 - Mediante alteración de la señalización celular. 2- Mediante influencia en las rutas metabólicas. 3 - Interacción directa con factores de transcripción para activar la expresión de los genes.

2.1 ÁCIDOS GRASOS ESENCIALES Y FACTORES DE TRANSCRIPCIÓN

Los ácidos grasos esenciales son aquellos que el cuerpo humano no es capaz de sintetizar, y por ello deben ser adquiridos a través de la dieta. Se trata principalmente de ácidos grasos poliinsaturados (PUFAs), lo cual significa que contienen dobles enlaces entre dos átomos de carbono a lo largo de la cadena de carbón (**Figura 3**). Concretamente presentan un doble enlace situado más cerca del extremo metilo terminal (opuesto al carboxilo, extremo omega) de la cadena del ácido graso, ya sea en el carbono 3 o en el carbono 6 (omega 3 u omega 6 respectivamente). Los mamíferos son incapaces de sintetizar ácidos grasos de este tipo, con enlaces dobles más allá de los

carbonos 9 y 10, contando desde el extremo carboxilo terminal de la cadena de ácido graso. Esta configuración molecular es la que confiere a los ácidos grasos esenciales sus propiedades saludables así como su potencial nutrigenómico. Estos ácidos grasos se encuentran en alimentos tales como el pescado azul, los frutos secos, las semillas y aceites vegetales.

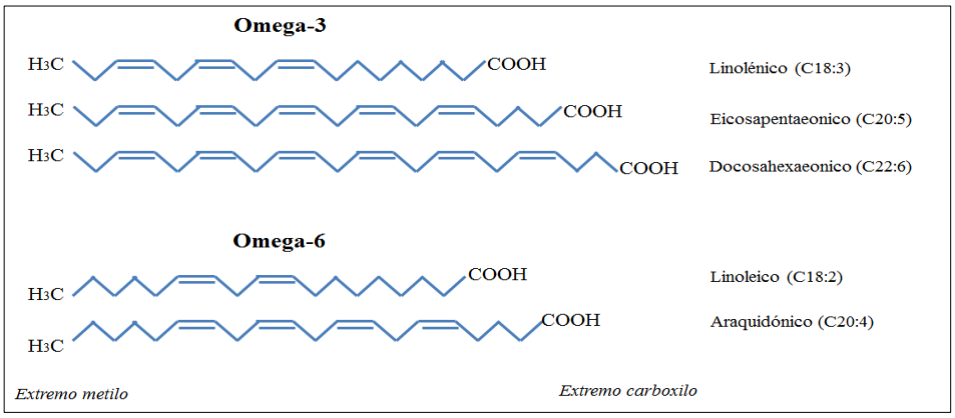


Figura 3: Estructura química de distintos ácidos grasos poliinsaturados.

Entre los efectos fisiológicos que provoca el consumo de ácidos grasos esenciales en la dieta se encuentra la reducción de los triglicéridos presentes en el plasma, el aumento de la beta oxidación de ácidos grasos y el aumento de la síntesis de fosfolípidos en el hígado (17). Así mismo se ha demostrado que su consumo mejora la presión arterial, la función endotelial y también reduce la inflamación (18). A nivel molecular, también tienen efectos en la estructura y plasticidad de las membranas celulares, y en las propiedades de los canales de iones. En general está bien demostrado que los PUFAs son potentes bioactivos, con un papel importante en el mantenimiento de la salud cardiovascular. Multitud de organismos nacionales e internacionales recomiendan un consumo mínimo de ácidos grasos esenciales.

Los ácidos grasos, en particular el ácido oleico (MUFA), los ácidos linoleicos y araquidónicos (PUFAs esenciales) han demostrado ser ligandos de muy alta afinidad para los receptores activados por proliferadores de peroxisomas alfa y gamma (PPARα y PPARγ) (**Figura 4**) (19). Estos receptores nucleares ejercen como factores de transcripción tras su activación, interactuando directamente con la cadena de ADN gracias a 2 dedos de zinc de la proteína que reconoce una secuencia de ADN específica, llamada elemento de respuesta a PPAR (PPRE), regulando así una gran diversidad de funciones biológicas. PPARα tiene un importante papel en el catabolismo de los lípidos y su homeostasis a nivel del hígado, mientras que por su parte PPARγ está implicado en el almacenamiento de los lípidos en el tejido adiposo. Además, la activación de ambos receptores PPAR inhibe la vía metabólica proinflamatoria del factor nuclear potenciador de las cadenas ligeras kappa de las células B activadas (NF-κB), con lo cual también poseen efectos antiinflamatorios (20).

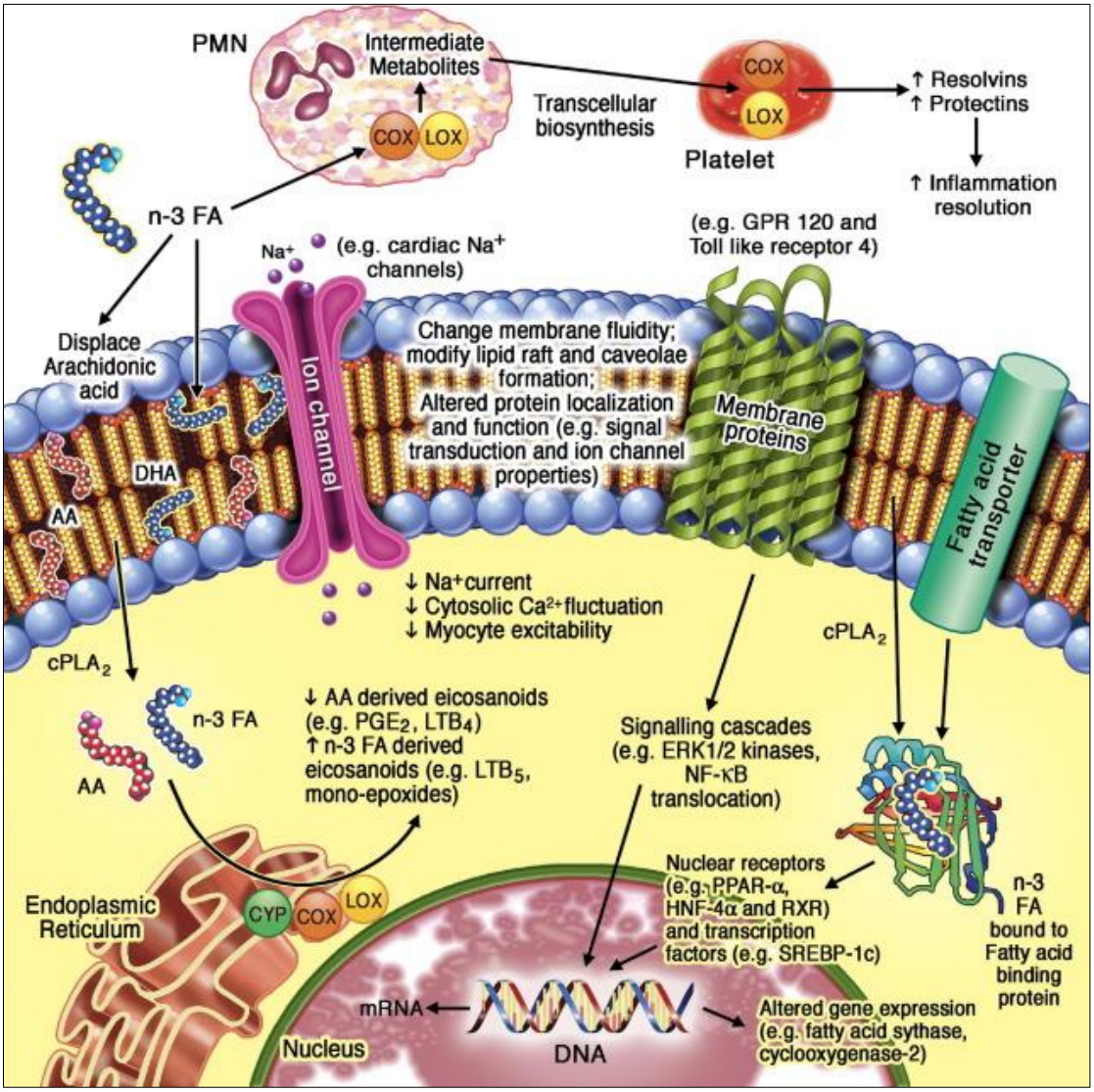


Figura 4: Visión general de los efectos moleculares de los PUFAs omega-3. Imagen extraída de la publicación original de acceso abierto (21). La incorporación de PUFAs omega-3 en la membrana celular altera su plasticidad, y éstos también interactúan con las proteínas de membrana y canales transportadores (parte central de la imagen). Una vez en el interior celular, se unen a receptores nucleares y factores de transcripción, para dirigirse al núcleo celular y alterar la expresión génica (parte inferior derecha de la imagen). Del mismo modo, los PUFAs omega-3 del citoplasma son convertidos a eicosanoides por las enzimas ciclo-oxigenasa (COX) y lipo-oxigenasa (LOX), así como proteínas de la familia citocromo P450 (CYP450), influyendo en importantes funciones biológicas como los procesos de inflamación (parte inferior izquierda de la imagen). En el exterior celular, metabolitos intermediarios como las resolvinas y las protectinas, derivados de las reacciones catalizadas por COX y LOX en al menos 2 tipos celulares distintos, serían importantes actores en procesos biológicos que reducirían la inflamación crónica en distintos modelos animales (parte superior de la imagen).

2.2 Señalización celular: ejemplo del cáncer

Los compuestos derivados de la dieta tienen un importante papel en multitud de vías metabólicas, y por ello pueden influir en la progresión de enfermedades como el cáncer. Muchos compuestos usados en alimentación, en su mayoría provenientes de plantas, han demostrado tener propiedades anticancerígenas, ya que son capaces de inhibir la proliferación celular y provocar

apoptosis *in vitro* en distintos tumores humanos, e *in vivo* en experimentos con ratones (22). Por ejemplo, estudios epidemiológicos han constatado que las poblaciones que consumen té verde regularmente presentan una menor incidencia de cáncer de mama y próstata a nivel mundial (23).

Estos efectos anticancerígenos podrían tener su origen en la capacidad de ciertos componentes de la dieta para modificar la transducción de señal a nivel celular. Es el caso de ciertos polifenoles presentes en el té verde, como el epigalocatequina-galato (EGCG). Se ha demostrado que este compuesto es capaz de inhibir la ruta de señalización PI3K (fosfoinositida-3-quinasa), activada tras la unión de multitud de factores de crecimiento (EGF, PDGF, VEGF) a receptores tirosina quinasa presentes en las membranas celulares. De esta manera se inhibe la consecuente cascada de señalización celular, evitando así la activación de la proteína quinasa B (Akt) por fosforilación. El resultado de esta inhibición es un aumento de la actividad transcripcional de los factores FOXO, una importante familia de factores de transcripción implicados en el control del ciclo celular, lo cual induce apoptosis deteniendo el ciclo celular (24). El EGCG también ha demostrado ser inhibidor, a través del mismo mecanismo PI3K/Akt, de la vía metabólica NF-κB, cuya activación está asociada con cánceres de mama, próstata, ovario y pulmón (25).

2.3 Efectos antiinflamatorios

La inflamación es una respuesta biológica de protección que se activa para hacer frente a lesiones celulares e infecciones por patógenos. Implica la liberación secuencial de distintas moléculas mediadoras de la inflamación, el reclutamiento y activación de células específicas del sistema inmune como linfocitos y macrófagos. De esta manera se activa una compleja red de intercomunicación celular en la zona inflamada a través de la liberación por los leucocitos de distintas citocinas proinflamatorias: el factor de necrosis tumoral alfa (TNFα), el interferón gamma (IFN-γ), y varias interleucinas (IL-1, IL-2, IL-6, IL-12, IL-15, IL-18, IL-22, IL-23) (26). Estas moléculas juegan un importante papel en la señalización celular, ya que pueden unirse a receptores específicos de las células y desatar así cascadas de señalización celular que modifican la expresión de los genes. La persistencia de los procesos inflamatorios se define como inflamación crónica y es perjudicial para la función fisiológica de los tejidos. De hecho la inflamación crónica está estrechamente relacionada con ECV, diabetes, enfermedades metabólicas y cáncer.

Alimentos ricos en bioactivos antiinflamatorios, tales como el ácido caféico (Yerba mate), el hidroxitirosol (aceite de oliva), la quercetina (flavonol de frutas y verduras), el licopeno (caroteno del tomate y sandía) o el α-tocoferol (té verde), han demostrado una gran capacidad para reducir la inflamación. Entre otros mecanismos, su actividad antiinflamatoria es debida a la capacidad de estos compuestos para reducir la expresión de genes de marcadores proinflamatorios tales como la

enzima ciclo-oxigenasa-2, codificada por el gen de la prostaglandina-endoperóxido sintasa 2 (PTGS2), el TNFα, así como de las óxido-nítrico sintasas (NOS1, NOS2, NOS3). Esta actividad puede ser explicada por la capacidad de estos compuestos para inhibir la vía metabólica del NF-κB. En efecto, esta vía también es muy importante en la regulación de la respuesta inmune y se encuentra activada crónicamente en enfermedades inflamatorias. Los compuestos antiinflamatorios derivados de la dieta, y particularmente los flavonoides, son capaces de inhibir en macrófagos la fosforilación del polipéptido alfa (IκBα) perteneciente al complejo proteico NF-κB (27). Esto impide la translocación, desde el citoplasma hasta el núcleo, de las cadenas ligeras kappa, impidiendo así la interacción de este complejo proteico con el ADN celular y la posterior cascada transcripcional proinflamatoria.

2.4 Efectos antioxidantes

La acumulación de especies reactivas de oxígeno en el interior de las células contribuye de manera sustancial al declive de las funciones fisiológicas del cuerpo humano. Algunas de estas especies son el superóxido (O_2-), el peróxido de hidrógeno (H_2O_2), el radical hidroxilo (• OH), el monóxido de nitrógeno (NO), y se producen de manera inevitable durante la respiración celular que tiene lugar en las mitocondrias, mediante el proceso metabólico de la fosforilación oxidativa. El estrés oxidativo que produce la acumulación de estas moléculas en las células afecta a la estructura de proteínas, lípidos y ADN. Se considera que las especies reactivas de oxígeno contribuyen de manera importante en la aparición de enfermedades como la diabetes, la ateroesclerosis, el cáncer e incluso algunas enfermedades neurodegenerativas (28).

Los polifenoles, un grupo de sustancias químicas encontradas en plantas, poseen importantes propiedades antioxidantes. Algunos de los compuesto fenólicos más estudiados obtenidos a través de la dieta son el resveratrol (uva), el hidroxitirosol (oliva), la quercetina (cebolla y brócoli) y la curcumina (cúrcuma común en India). A nivel cardiovascular, se ha demostrado que compuestos como la quercetina son capaces de mantener estables los niveles de un importante vasodilatador en los vasos sanguíneos, el óxido nítrico. Esto ocurre gracias a la capacidad de la quercetina para capturar moléculas de superóxido, mejorando de esta manera la hipertensión y disminuyendo los niveles de lipoproteínas de baja densidad (LDL) oxidadas en la sangre (29,30). Estudios en hepatocitos también han demostrado la capacidad de la quercetina para modular gran variedad de genes con capacidad antioxidante (31-33).

Otros fitoquímicos como los carotenos (especialmente el beta-caroteno) y los tocoferoles han demostrado su actividad antioxidante debido a la capacidad de estas moléculas para capturar radicales libres (34).

Del mismo modo el resveratrol (RVT) ha demostrado tener importantes propiedades antioxidantes *in vitro*, aunque también se han descrito muchas otras actividades biológicas. Éstas podrían explicarse por su capacidad de interacción con una importante enzima histona deacetilasa, la Sirtuina 1 (Sirt1). Esta proteína es capaz de fijar numerosos sustratos y es muy importante en la regulación del metabolismo. El RVT puede incrementar la actividad de Sirt1 hasta en 20 veces (35). El mecanismo de acción se basa en la capacidad de la molécula de RVT para unirse a la extremidad nitrogenada sin actividad catalítica de la proteína Sirt1, provocando así un cambio de conformación en la enzima. El cambio provocado aumenta la actividad de Sirt1 ya que su afinidad para fijar sustratos acetilados se ve incrementada, afectando así a la transducción de señal por el aumento de la fosforilación de la proteína quinasa activada por adenosina monofosfato (AMPK), reduciendo también marcadores de estrés oxidativo (36).

Sirt1 es una importante diana farmacológica de enfermedades metabólicas y cardiovasculares, y se han invertido importantes recursos en la investigación de activadores de esta enzima. El mejor ejemplo lo representa la antigua empresa biotecnológica Sirtris Pharmaceuticals, cuya actividad se concentraba únicamente en encontrar activadores de Sirt1. Además, existen en el mercado nutracéuticos como Longevinex®, compuestos de RVT modificado para una mayor biodisponibilidad, y que entre otros efectos reivindican una mayor protección cardiovascular.

2.5 Potencial de los micro ARN's

Los micro ARN's (miARN) son moléculas cortas de ARN (entre 18 y 25 nucleótidos) que actúan como reguladores post-transcripcionales, reprimiendo simultáneamente la traducción en proteína de múltiples ácidos ribonucleicos mensajeros (ARNm) de potenciales genes diana, afectando de esta manera a vías metabólicas específicas. Actualmente hay descritas más de 2.656 secuencias de miARN's humanos en su forma madura en miRBase (versión 22.1, Octubre 2018), la principal base de datos sobre esta temática (37).

La biogénesis de los miARN's empieza a partir de la transcripción de una región genómica determinada en una molécula de pri-miARN. Tras la transcripción del pri-miARN en el núcleo celular por la ARN polimerasa II, la ribonucleasa Drosha se encarga de hidrolizar la parte de la molécula de ARN que no se encuentra unida por complementariedad. La molécula así obtenida es un pre-miARN de entre 60 y 70 nucleótidos, con una estructura en forma de horquilla, la cual es transportada hasta el citoplasma celular a través de la proteína Exportina 5. Posteriormente la proteína Dicer se encarga de realizar una escisión en el pre-miARN, obteniendo así una molécula de miARN de doble cadena, con una longitud de entre 18 y 25 pares de bases. Esta molécula dará lugar a 2 moléculas de miARN maduras (5p y 3p), las cuales podrán ser integradas separadamente en el complejo RISC para realizar

así la tarea de represión de la traducción de determinados ARNm, seleccionados en base a la complementariedad entre la región semilla ("seed") del miARN maduro, situada entre los nucleótidos 2-7 desde el extremo 5', y las secuencias del ARNm, generalmente situadas en la región 3' de la región no traducida del ARNm. Si bien la función canónica de los miARN's se centra en la represión de la traducción de secuencias específicas mediante interacción directa con las moléculas de ARNm diana (38), existe evidencia de que también podrían actuar como activadores de la expresión génica de manera indirecta, interactuando con las secuencias promotoras de los genes (39).

2.5.1 Regulación de micro ARN's endógenos:

Se ha demostrado que componentes de la dieta son capaces de regular determinados miARN's con importantes consecuencias fisiológicas. En uno de los primeros trabajos de investigación sobre esta temática, se observó que los niveles del miARN miR-33 presentaban una fuerte correlación negativa con las concentraciones de colesterol intracelular, sugiriendo una regulación de la expresión de este miARN por los niveles de colesterol ingeridos a través de la dieta (40). En el mismo trabajo se demostró que miR-33 es capaz de reprimir la expresión de genes implicados en el flujo de colesterol celular en hepatocitos humanos y de rata, como algunas proteínas transportadoras ABC dependientes de trifosfato de adenosina (ABCA1 y ABCG1).

Determinados miARN's también responden a la presencia de ácidos grasos en la dieta, como el ácido graso docosahexaenoico (DHA), y son capaces de regular la expresión de genes implicados en el metabolismo lipídico en células de enterocito humano (41). Además, mecanismos epigenéticos influenciados por la expresión de algunos miARN's podrían explicar la conexión entre dieta y salud mental. Autores han sugerido que la dieta mediterránea tendría un efecto potencialmente neuroprotector (42). Estudios de dietas suplementadas con concentrados fosfolipídicos de aceite de krill y suero de leche han demostrado regular la expresión de gran cantidad de miARN's en el hipocampo cerebral de ratas (43). En efecto el análisis funcional de los miARN´s regulados reveló una estrecha asociación con mecanismos relacionados con el sistema nervioso, incluyendo procesos de neurogénesis y plasticidad sináptica.

2.5.2 Presencia y actividad de micro ARN's exógenos

Más recientemente, investigaciones han detectado niveles de miARN's provenientes del arroz circulando en la sangre del organismo humano. Esta observación resulta controvertida para la comunidad científica, ya que significaría que material genético ajeno al organismo humano podría estar influyendo en la expresión de nuestros genes (44). La posibilidad de absorber miARN's

exógenos a través del intestino humano podría abrir otra puerta para explicar los beneficios generales a nivel cardiovascular de dietas con abundantes vegetales o incluso de la dieta mediterránea. Si bien los miARN's exógenos deberían ser degradados a través de su paso por el aparato intestinal, se ha demostrado que en las plantas comestibles existen unas nanopartículas, similares a los exosomas de mamíferos, las cuales podrían facilitar la transferencia de material genético entre especies. Estas nanopartículas, o vesículas extracelulares, contendrían miARN's, proteínas y otros lípidos, y serían absorbidas por los macrófagos del intestino, preservando así la estabilidad de estas moléculas y permitiendo ejercer un efecto biológico (45).

2.6 Efectos en la metilación del ADN

La expresión de los genes se ve afectada por cambios estructurales en el ADN más allá de las modificaciones a nivel de la secuencia genómica. Cada tipo celular posee su propio patrón de metilación de ADN, permitiendo así la expresión y el silenciamiento específico de los genes.

El efecto de la dieta en el patrón de metilación del ADN todavía es un terreno de estudio con un amplio potencial. En uno de los pocos estudios al respecto, una intervención de 5 años de duración con 36 participantes adheridos al consumo de la dieta mediterránea, se observaron cambios en la metilación de las células mononucleares de sangre periférica de los participantes. En línea con las propiedades antiinflamatorias observadas con la multitud de fitoquímicos que componen la dieta mediterránea, estos cambios correspondieron a genes asociados a procesos de inflamación e inmunocompetencia: la interleucina 4 inducida 1 (IL4I1), el receptor de leptina (LEPR), el coactivador 1 beta de PPARG (PPARGC1B), e importantes quinasas como la proteína quinasa 2 activada por MAP quinasas (MAPKAPK2), entre otros genes (46).

3. Herramientas para el análisis del nutrigenoma

Durante los últimos 20 años, las herramientas de tipo Ómicas han contribuido de manera sustancial al conocimiento de los mecanismos moleculares de las enfermedades metabólicas y a la identificación de biomarcadores. La integración de estas herramientas en la investigación nutrigenómica, en combinación con disciplinas como la bioinformática, es un proceso necesario y fundamental para profundizar en el conocimiento de la regulación del nutrigenoma, término con el cual se refiere a la capacidad de los alimentos y sus compuestos bioactivos para regular la expresión de los elementos que componen el genoma.

En concreto, la transcriptómica permite cuantificar simultáneamente el nivel de expresión de los elementos que componen el genoma, ya sean genes, largos ARNm no codificantes o miARN's, con el fin de detectar y poder explicar a nivel molecular las diferencias fenotípicas entre condiciones de referencia y experimentales. Existen distintas técnicas y plataformas para realizar estudios de transcriptómica, y cada cual requiere de métodos distintos para el análisis de los datos obtenidos.

Junto a todas las tecnologías que permiten generar datos biológicos de forma masiva, también es importante el componente analítico de los datos obtenidos. Con esta necesidad se creó el proyecto Bioconductor, basado en el lenguaje de programación R, de código abierto y que proporciona multitud de métodos para el análisis de datos en Genómica. El lenguaje de programación R ya contaba con la implementación de herramientas de análisis estadístico y visualización de datos, y además ofrece una interfaz relativamente sencilla y rápida para prototipar nuevos métodos computacionales. Efectivamente la gran cantidad de datos generados por las tecnologías de tipo Ómicas requiere de nuevos métodos computacionales y modelos estadísticos para su análisis (47).

Como en todo análisis estadístico de datos, el diseño experimental es muy importante para la obtención de resultados robustos. El número de réplicas técnicas y biológicas, el nivel de correlación de datos intragrupales y la detección de valores extremos son pasos necesarios para la obtención de resultados fiables.

A continuación se exponen las distintas plataformas de transcriptómica utilizadas durante la realización de esta tesis doctoral para el estudio de la regulación de la expresión de los genes por los alimentos y sus compuestos bioactivos, así como su correspondiente método utilizado para el análisis de los datos.

3.1 ARRAYS DE EXPRESIÓN

Con el objetivo de construir una base de datos con los resultados de experimentos de nutrigenómica, en este trabajo se han coleccionado y analizado datos públicos generados con miroarrays de un color (análisis de una única muestra hibridada a un solo fluoróforo), microarrays de dos colores (análisis simultáneo de dos muestras hibridadas a dos fluoróforos) y arrays de expresión de tipo "beadchip".

3.1.1 MICROARRAYS

La tecnología de los microarrays se desarrolló a principios del año 2000. Esta tecnología ha progresado rápidamente gracias a los nuevos métodos de producción y detección de señales de fluorescencia, y también gracias a la disponibilidad de la secuencia completa de múltiples genomas,

y en especial el primer borrador del genoma humano, obtenido en abril del año 2000. Esta herramienta permite cuantificar el nivel de expresión de genes mediante el análisis de la abundancia de ARNm. Se trata de microchips capaces de analizar simultáneamente la expresión de todos los genes conocidos que componen el genoma humano, produciendo así una gran cantidad de datos. Cada una de las miles de celdas de estos microchips contiene impresas cortas secuencias específicas de ADN, oligonucleótidos de distinta longitud en función de la plataforma utilizada. Un chip de microarray puede llegar a contener hasta 250.000 oligonucleótidos por centímetro cuadrado. Estos oligonucleótidos actúan como sondas de detección, y generalmente se emplean distintas sondas para la detección de un mismo gen, con el objetivo de obtener datos de expresión fiables.

Su principio está basado en la hibridación por complementariedad de 2 secuencias de nucleótidos. Tras obtener la muestra de ARNm a analizar, ésta es retro transcrita a ácido desoxirribonucleico complementario (ADNc), amplificada por la reacción en cadena de la polimerasa (PCR) y marcada mediante uno de los fluoróforos más utilizados; las cianinas 3 y 5 (Cy3, Cy5). Después se mide el nivel de hibridación entre las secuencias presentes en el microarray y el ADNc de la muestra a analizar (**Figura 5**). La intensidad de fluorescencia detectada en una celda del microchip correlaciona directamente con la cantidad de ARNm hibridado, y corresponde al nivel de expresión de un gen de secuencia conocida. En el caso de los microarrays de dos colores, se analizan dos muestras marcadas con ambos fluoróforos en el mismo chip, resultando en una hibridación competitiva, obteniendo tras el escaneo una imagen con una mezcla de ambos colores y con intensidades variadas (**Figura 6**). Tras escanear el microarray (excitación del fluoróforo y captura de fluorescencia), se obtiene una imagen que refleja el nivel de fluorescencia presente en cada una de las múltiples celdas del microarray. Finalmente, se analiza la imagen obtenida mediante programas informáticos, obteniendo un fichero de formato texto con la cuantificación de la señal lumínica generada en cada celda del microarray; los llamados datos crudos.

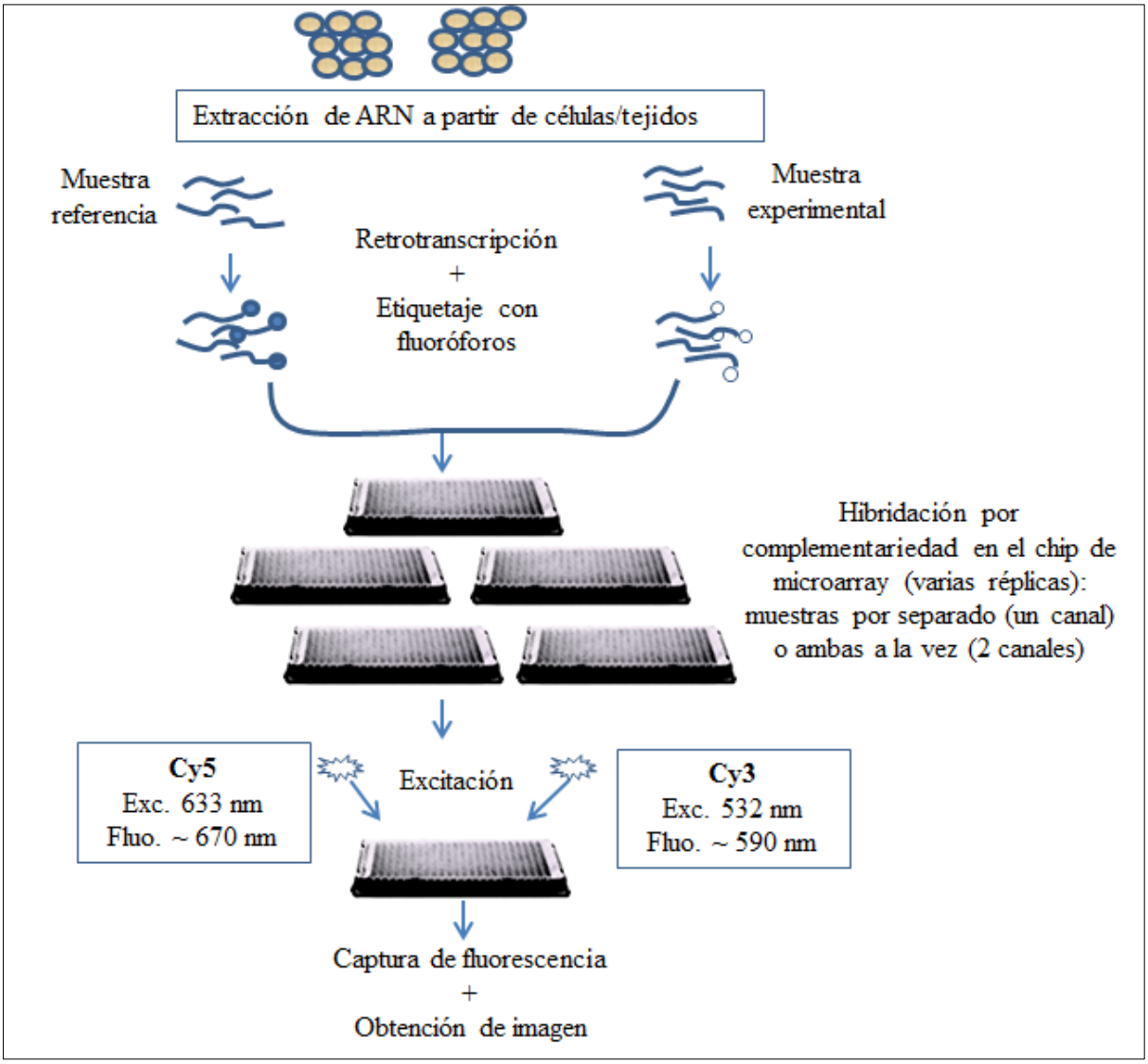


Figura 5: Esquema de las etapas necesarias para la realización de un experimento de expresión diferencial con microarrays.

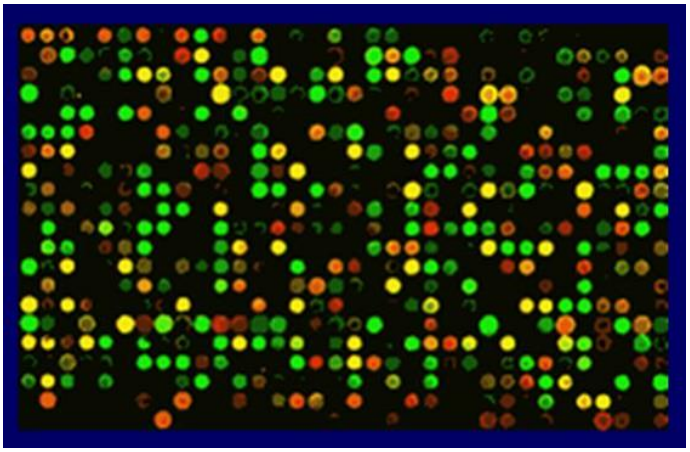


Figura 6: Imagen obtenida tras el escaneo de un microarray analizado por hibridación competitiva. Se superponen las fluorescencias obtenidas tras excitar ambos fluoróforos por separado.

Uno de los prerrequisitos para el uso de esta tecnología, basada en la hibridación por complementariedad de ácidos nucleicos, es que las secuencias impresas en el microchip deben de ser conocidas, para así saber qué es lo que se va a detectar. Una limitación importante es la hibridación cruzada, que se produce cuando se analizan cortas secuencias de ADN muy similares. De la misma manera, debido a su principio de detección, resulta difícil medir con precisión la expresión de ARNm poco abundante. También es importante considerar que la señal fluorescente que mide el nivel de hibridación sonda/ARNm no siempre es linealmente proporcional a la concentración de este último. Efectivamente cuando un ARNm se encuentra presente en altas concentraciones, puede saturar las sondas que lo detectan, y al contrario, en muy bajas concentraciones puede que no se produzca hibridación alguna con la sonda. De este modo la señal detectada es lineal únicamente dentro de un determinado rango de concentraciones (48).

Los microarrays de un único color se comercializan principalmente por la empresa Affymetrix, mientras que los microarrays de dos colores son propios de la empresa Agilent.

3.1.1.1 MICROARRAYS DE AFFYMETRIX

En este caso las distintas secuencias de oligonucleótidos presentes en el chip son impresas directamente en un proceso de síntesis *in-situ* sobre un soporte de vidrio, mediante la técnica de fotolitografía. Los oligonucleótidos que ejercen como sondas de detección ("probes") suelen tener una longitud de 25 nucleótidos. Generalmente emplean múltiples sondas diferentes para cada gen a detectar, aumentando así la especificidad de la detección, lo cual da lugar a los denominados "probesets". Sus chips cuentan con una alta densidad de sondas, de hasta 500.000 por microchip. Este fabricante dispone principalmente de dos distintos microchips de microarrays que difieren por el modo de hibridación:

- modelos 3'IVT: las sondas del microarray detectan zonas del extremo 3' del ARNm, muchas veces en la zona no traducida (3'UTR). Estas zonas presentan una amplia diversidad de secuencia y distinguen muy bien los distintos transcritos presentes en la célula. Una de las plataformas de este tipo más utilizadas es la Human Genome U133 Plus 2.0 Array, la cual permite analizar hasta 54.681 elementos del genoma, principalmente transcritos codificantes.

- modelos Whole Transcript: las sondas del microarray se distribuyen a lo largo de la secuencia completa de un gen, incluyendo los distintos exones. Proporcionan una señal más fiable, además de la posibilidad de detectar distintas isoformas de transcritos correspondientes a un mismo gen. Una de las plataformas de este tipo más utilizadas es la

Human Gene 2.0 ST Array (HuGene-2_0-st), la cual permite analizar hasta 53.981 elementos del genoma incluyendo genes y ARNm no codificantes.

3.1.1.2 MICROARRAYS DE AGILENT

La empresa Agilent emergió como una Spin-off de Hewlett-Packard. Las secuencias de oligonucleótidos también se imprimen directamente sobre el soporte de vidrio del microchip mediante un proceso de síntesis *in-situ*, utilizando su tecnología propia de inyección de tinta (tecnología SurePrint). Se sintetizan las distintas secuencias base a base, repitiendo capas de impresión utilizando el método químico de los fosforamiditos. De esta manera se pueden sintetizar sondas de hasta 60 nucleótidos, mejorando así la especificidad de detección, pero también reduciendo el número de genes a analizar. Además estas plataformas cuentan con gran cantidad de sondas correspondientes a largos transcritos no codificantes. A diferencia de los "probesets" característicos de Affymetrix, el número de sondas para detectar un mismo transcrito es menor debido a la mayor longitud de estas. Uno de los modelos de microarrays más conocido es el SurePrint G3 Human Gene Expression 8x60K, permitiendo analizar hasta 62.976 distintos elementos del genoma incluyendo genes y ARNm no codificantes.

3.1.1.3 MICROARRAYS FABRICADOS A MEDIDA

Existe la posibilidad de pedir a estos fabricantes la fabricación de microarrays con sondas específicas a gusto del investigador. En estos casos, el fabricante Agilent dispone de microarrays de ADNc ("cDNA microarrays"). Estos microarrays contienen impresas largas secuencias de ADN complementario, y a diferencia de sus chips clásicos, esta vez las secuencias son sintetizadas a parte e impresas posteriormente de manera mecánica en el microchip. En el caso del fabricante Affymetrix, también dispone de arrays específicos. Un buen ejemplo es el array NuGo, compuesto por 24.000 "probesets" que examinan genes relevantes a estudios de nutrigenómica.

3.1.2 ARRAYS DE EXPRESIÓN "BEADCHIP" DE ILLUMINA

Estos arrays de expresión aparecieron en el mercado posteriormente, y solucionan varias limitaciones presentes en los chips de microarrays. Permiten analizar entre 6 y 8 muestras en el mismo array de expresión de manera simultánea, además de requerir una menor cantidad de material genético, lo cual evita la etapa de amplificación de ADNc. Gracias a la tecnología propia del fabricante Illumina, se fijan oligonucleótidos de 50 bases en micro esferas, las cuales se introducen en el array para su reorganización espontánea en cada uno de los pocillos disponibles. Cada esfera contiene cientos de miles de sondas de secuencia única, ligadas de manera covalente, obteniendo así una alta redundancia para cada secuencia representada en el array. Actualmente esta plataforma

de arrays de expresión es la de mayor densidad de sondas por milímetro, y resulta económicamente ventajosa en comparación con los microarrays clásicos. Por ejemplo su plataforma HumanRef-8 v2 beadchip permite analizar 8 muestras de manera simultánea, y detecta hasta 22.185 distintos elementos del genoma, principalmente ARNm.

3.1.3 Análisis de datos de arrays de expresión

Una vez procesada la imagen obtenida tras la lectura de los niveles de fluorescencia, y antes de iniciar un análisis de expresión diferencial, es necesario realizar tres etapas esenciales: ajuste del ruido de fondo, normalización y resumen de los datos. Tras ello es recomendable anotar los identificadores de las sondas con el nombre del gen correspondiente.

La normalización de los datos crudos es diferente dependiendo de si se trata de un microarray de uno o de dos colores. En el caso de los microarrays de un sólo color analizados en este trabajo, se ha empleado el algoritmo de "robust multi-array average" (RMA) (49). El algoritmo RMA crea una matriz de datos de expresión a partir de los datos crudos; elimina el ruido de fondo, transforma los datos a escala logarítmica en base 2, y realiza la normalización de los datos, igualando la distribución de las intensidades obtenidas tras el escaneo de los distintos arrays de expresión incluidos en el experimento. Además también es capaz de resumir los datos de las distintas sondas de un "probeset", obteniendo así un único valor de expresión para el gen analizado en dicho "probeset". La implementación de este algoritmo en el proyecto Bioconductor puede encontrarse en las librerías "affy" y "oligo". En los casos de los microarrays de dos colores analizados en este trabajo, así como los datos generados mediante los arrays de expresión tipo Beadchip, se han utilizado directamente las matrices de datos normalizadas por el software propio de Agilent e Illumina. Tras este proceso, se anotan las sondas analizadas en función de la plataforma utilizada, identificando así cada gen analizado.

Para el análisis de expresión diferencial de genes, dentro del proyecto Bioconductor existen varias librerías que implementan distintos algoritmos de análisis. Las más importantes son las librerías Limma (50) y "Significance Analysis of Microarrays" (SAM) (51); ambas asumen una distribución distinta para los datos, y a diferencia de Limma, SAM utiliza métodos estadísticos no paramétricos para el análisis. En este trabajo se ha empleado la librería Limma para el análisis de datos de los microarrays y arrays de expresión.

Limma es la librería que implementa métodos estadísticos paramétricos para el análisis de datos de expresión diferencial generados por microarrays. El principio de análisis se basa en modelizar de manera lineal los datos de expresión de cada gen. Los datos de expresión pueden ser

proporciones de intensidades (caso de dos canales) o intensidades de un solo fluoróforo. Esta información debe encontrarse en escala logarítmica con el fin de reducir la variabilidad de los datos. En el método Limma, los datos se modelan mediante un modelo lineal:

$$E[\ y_j\ ] = X\alpha_j$$

Donde y_j contiene los datos de expresión para el gen j, X es la matriz de diseño, y α_j es un vector de coeficientes.

Para obtener la significancia estadística del contraste realizado, es necesario realizar un test de hipótesis y ajustar el valor *p* debido al testeo simultáneo, dado que se están analizando miles de genes al mismo tiempo. El test estadístico utilizado es una prueba *t* moderada para cada gen analizado, equivalente a una simple prueba *t*, habiendo moderado antes los errores estándar entre todos los genes analizados.

Los resultados obtenidos gracias al algoritmo implementado en la librería Limma incluyen el nivel de expresión diferencial ("fold change") para cada gen tras el experimento, el nivel medio de expresión en todos los arrays del experimento, el valor del estadístico *t* y los valores *p* (**Tabla 2**). El paso final del proceso analítico trata de corregir los valores *p* obtenidos, con el fin de reducir los falsos positivos que se producen debido al testeo simultáneo. Efectivamente la prueba *t* moderada se realiza con un umbral de confianza al 5%, con lo cual existe un 5% de probabilidades de rechazar incorrectamente una hipótesis nula. En el caso de los experimentos de transcriptómica, dónde se testean miles de hipótesis de manera simultánea, la probabilidad de rechazar incorrectamente la hipótesis nula crece en función del número de genes analizados en la plataforma. Para realizar la corrección, los métodos más empleados en bioinformática son los ajustes de "False Discovery Rate" (FDR) (52), de Bonferroni y de Holm. Gracias a esta corrección se obtiene una lista de genes diferencialmente expresados en distintos niveles y significancias estadísticas.

Tabla 2: Ejemplo de un fichero de resultados obtenido tras el análisis de expresión diferencial de un experimento de microarrays usando la librería "Limma", y tras anotación de las sondas con su correspondiente nombre de gen.

Símbolo	Log2FC	AveExpr	*t*	Valor *p*	Valor *p* ajustado	B
DKK1	-3.581861918	6.427131469	-44.80666403	4.07E-13	2.23E-08	17.80662222
RSAD2	3.346348362	7.688660157	38.83631617	1.76E-12	3.95E-08	17.08098405
CTGF	-3.272189427	11.34070516	-37.27586721	2.67E-12	3.95E-08	16.85224714
LINC01013	-3.142682824	6.75476514	-36.98727445	2.89E-12	3.95E-08	16.80785912
ANGPT2	-3.339681793	8.501086957	-33.53744076	7.83E-12	8.57E-08	16.22056452
IFIT1	3.128256929	9.157998741	31.8851535	1.31E-11	1.01E-07	15.8974218

Símbolo	Log2FC	AveExpr	*t*	Valor *p*	Valor *p* ajustado	B
OAS1	2.568237233	7.785006835	31.32099395	1.57E-11	1.01E-07	15.78003164
ANGPT2	-3.172955655	8.153584625	-31.29561044	1.58E-11	1.01E-07	15.77466091
OAS1	2.655684768	7.598224718	28.82280578	3.66E-11	1.73E-07	15.21192625
CMPK2	2.940011338	9.432141052	28.6603999	3.87E-11	1.73E-07	15.1720473
IFIT3	2.751580601	10.24492522	28.48912223	4.12E-11	1.73E-07	15.12957262
IFIT2	3.10449648	8.222017967	28.16361328	4.62E-11	1.81E-07	15.04765049
RAB11FIP1	-2.526155036	5.623301882	-26.26242972	9.40E-11	3.21E-07	14.5357897
OASL	2.763093782	7.085678894	25.94617737	1.06E-10	3.42E-07	14.44474584
APLN	-3.152922101	7.239307168	-25.2190573	1.42E-10	4.30E-07	14.22853941
MX1	2.237919476	9.656786447	24.5184146	1.88E-10	5.42E-07	14.01070696
CYP1A1	2.042447691	6.435475783	23.66365983	2.70E-10	7.10E-07	13.73154366
CXCR4	-2.287252621	5.726773413	-23.51755614	2.87E-10	7.10E-07	13.68228124
TNFSF13B	2.972856249	5.751271898	23.49183837	2.90E-10	7.10E-07	13.67356187
CCL5	2.372237473	8.622597677	23.42518808	2.99E-10	7.10E-07	13.65089738
SGK1	-1.892084262	9.541865001	-23.29246198	3.16E-10	7.21E-07	13.60547258
MCTP1	-2.125294787	8.28146532	-23.09580744	3.45E-10	7.53E-07	13.53744808
KYNU	2.133740225	6.93312368	22.92451559	3.71E-10	7.81E-07	13.47748585
TGFB2	-2.042913104	6.537906186	-22.79578832	3.93E-10	7.96E-07	13.43198205

3.2 SECUENCIACIÓN DE NUEVA GENERACIÓN

La secuenciación de nueva generación (NGS) se refiere a las nuevas tecnologías de secuenciación de ADN que emergieron a principios del año 2000 y revolucionaron la investigación genómica. El principio utilizado por estas nuevas tecnologías y la clásica secuenciación por el método de Sanger son similares: la secuenciación por síntesis. Ambos métodos están basados en la adición secuencial de nucleótidos fluorescentes por una enzima ADN polimerasa en base a la complementariedad con la molécula a secuenciar, y la posterior captura de señal fluorescente emitida en cada incorporación, la cual será diferente en función del nucleótido incorporado.

Las plataformas de NGS son capaces de secuenciar múltiples fragmentos de ADN de manera simultánea, generando así un gran volumen de datos en poco tiempo. Una variación técnica para la detección de nucleótidos incorporados utilizada en NGS, e implementada en los secuenciadores 454 de Roche, es la pirosecuenciación, basada en la detección del anión pirofosfato liberado tras la incorporación de cada nucleótido. En la actualidad ya es posible secuenciar un genoma humano en pocos días. Sin embargo el análisis de la multitud de datos generados por las plataformas de NGS representa un cuello de botella. La técnica de análisis del transcriptoma utilizando métodos de NGS adopta el término de RNA-Seq.

3.2.1 Principio de RNA-Seq

El principio de la técnica de RNA-Seq está basado en retrotranscribir y fragmentar de manera aleatoria la muestra de ARNm a analizar, para luego alinear los fragmentos obtenidos contra una base de datos de secuencias de referencia, utilizando distintas herramientas computacionales. El resultado de este proceso es una tabla con el recuento de la cantidad de lecturas de secuenciación correctamente mapeadas a posiciones genómicas, la cual puede contener del orden de millones de lecturas mapeadas. De esta manera, a diferencia del análisis de expresión con arrays, es posible obtener una cuantificación extremadamente precisa y altamente reproducible del nivel de expresión de los genes, evitando la etapa de transformar en un número la intensidad de una señal fluorescente. La técnica de RNA-Seq es considerada una tecnología digital de análisis de expresión, al contrario de los arrays de expresión, que son tecnologías analógicas. Del mismo modo, en RNA-Seq tampoco es necesario conocer previamente la secuencia a analizar, abriendo así la puerta a realizar nuevos descubrimientos. Esta técnica permite detectar con precisión distintas isoformas de ARNm y miARN, las cuales están implicadas en numerosas patologías humanas. Del mismo modo, gracias a esta técnica fue posible demostrar claramente que hasta tres cuartas partes del genoma humano podían ser transcritas en moléculas de ARN de distintas longitudes, sugiriendo así la necesidad de redefinir el concepto original de gen (53).

La primera etapa de cualquier proceso de RNA-Seq consiste en la preparación de la librería a secuenciar. La calidad de esta etapa es crítica para la obtención de buenos resultados a posteriori. Para ello es necesario realizar previamente la reacción de retrotranscripción del ARNm, fragmentar, y proceder a la ligación de una misma secuencia de ADN conocida en ambas extremidades de cada molécula (los adaptadores). Posteriormente, y con el fin de aumentar la señal generada, las secuencias fragmentadas de ADNc diana obtenidas, o insertos, son amplificadas por la reacción de la polimerasa en cadena (PCR), obteniendo así clústeres de moléculas idénticas a secuenciar. Una vez finalizado este proceso, es posible realizar la secuenciación de las moléculas, cuya estrategia varía dependiendo de la plataforma de NGS utilizada.

Existen tres parámetros importantes a tener en cuenta antes de realizar un proceso de secuenciación:

-tipo de carrera: durante la secuenciación pueden realizarse lecturas únicas o pareadas (**Figura 7**). Las lecturas únicas consisten en secuenciar un fragmento de ADNc a partir de un único extremo, mientras que con las lecturas pareadas los fragmentos son secuenciados primero desde un extremo y luego desde el extremo opuesto. Si bien la secuenciación por lectura única resulta suficiente para analizar estudios de RNA-Seq, la secuenciación por lecturas pareadas resulta más

ventajosa; permite un alineamiento más preciso con las secuencias de referencia, aumentando así la calidad del mapeo final. Además, y dependiendo de la aplicación de la secuenciación, las lecturas pareadas facilitan la detección de nuevos transcritos, reordenamientos genómicos, fusiones de genes y secuencias repetidas.

-longitud de lectura: depende de la plataforma utilizada y de los métodos de secuenciación comentados anteriormente. En general se asume que cuanto más larga sea una lectura, la probabilidad de que ésta se alinee sin ambigüedades con la secuencia de referencia será mayor, obteniendo así valores de mapeo óptimos. Sin embargo es importante considerar el coste económico de la secuenciación, ya que para obtener lecturas más largas se requieren más ciclos de secuenciación para la incorporación de nucleótidos, incrementando el coste económico del proceso. Además la elección de este parámetro depende del objetivo de la secuenciación. En la práctica la longitud de las lecturas generadas para experimentos de RNA-Seq varían entre 75 y 300 nucleótidos, mientras que se prefieren lecturas más largas para experimentos de re-secuenciación de ADN o secuenciación *de-novo*. Un caso especial es la secuenciación de miARN en su forma madura, donde las lecturas no exceden los 30 nucleótidos.

-profundidad de cobertura: dado que las carreras de secuenciación generan lecturas que se distribuyen de manera aleatoria e independiente a lo largo del ADN secuenciado, el objetivo es que todas las bases del material genético de la muestra sean secuenciadas varias veces. De este modo la profundidad de cobertura se refiere a la media del número de veces en que una base será leída durante la secuenciación. Cuanto mayor sea este parámetro, mayor será la calidad de los datos, ya que una base específica habrá sido secuenciada varias veces y no habrá lugar a dudas sobre posibles errores de secuenciación. La elección de una determinada profundidad de cobertura depende la aplicación que se va a dar al proceso de secuenciación. Para experimentos de RNA-Seq se considera suficiente una profundidad de cobertura de entre 10 y 25.

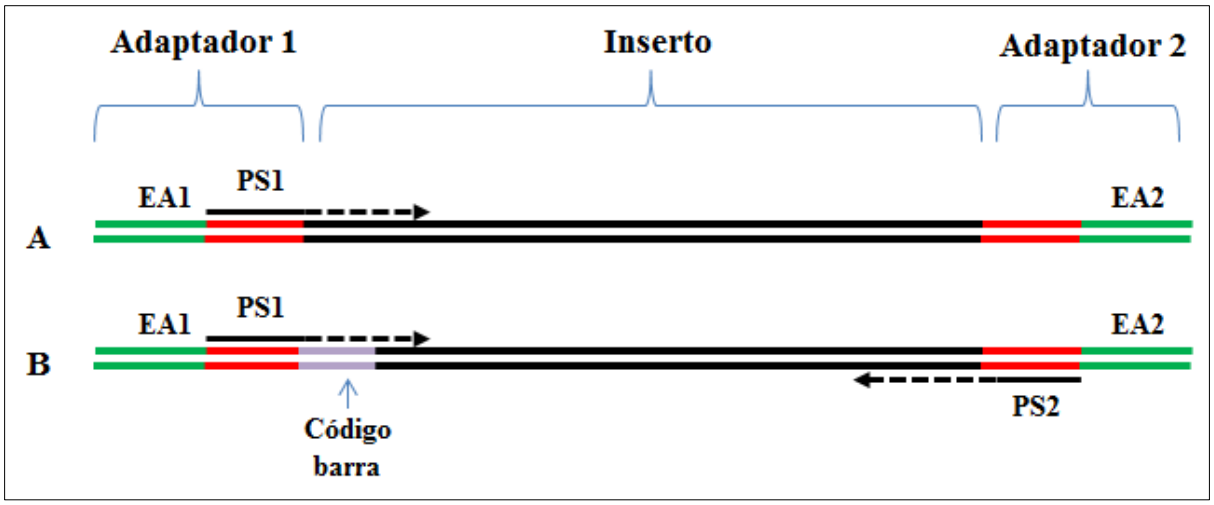


Figura 7: Esquema de una molécula a secuenciar resultante de la etapa de preparación de la librería. Ambos adaptadores contienen dos elementos funcionales: una secuencia correspondiente al elemento de amplificación (EA), y otra secuencia correspondiente a la sonda que inicia la secuenciación por síntesis (PS). A - Configuración de una librería para secuenciación de lectura única (una única sonda de secuenciación PS1). B - Configuración de una librería para secuenciación de lectura pareada (dos sondas de secuenciación PS1 y PS2), incluyendo una secuencia "Código de barra" para permitir la secuenciación de múltiples muestras en una misma carrera.

Últimamente se han realizado importantes avances en secuenciación de molécula única, proceso por el cual se obtienen resultados de secuenciación más fiables, debido principalmente a los errores aleatorios que puede introducir la enzima ADN polimerasa durante la etapa de amplificación a la hora de la construcción de los clústeres de moléculas a secuenciar. Del mismo modo se han desarrollado equipos específicos para la preparación de las librerías y que tienen en cuenta la hebra de ADN secuenciada, si bien esto no es crítico en la secuenciación de genomas/transcriptomas eucariotas, debido al escaso solapamiento de las regiones transcritas del ADN. Sin embargo, actualmente la tecnología más puntera en secuenciación la proporciona la empresa de biotecnología Oxford Nanopore, que desarrolla secuenciadores portátiles de reducido tamaño, y ofrecen la posibilidad de secuenciar moléculas únicas de manera electrónica, evitando así cualquier uso de fluorescencia inherente de la secuenciación por síntesis.

3.2.2 Análisis de datos

Como paso previo es recomendable realizar una etapa de control de calidad de los datos obtenidos por RNA-Seq. Esta etapa comprende principalmente la eliminación de las lecturas de baja calidad. La eliminación de los adaptadores utilizados durante la secuenciación no es un paso necesario cuando las lecturas son relativamente largas, pero sí es importante si se trabaja con secuencias de en torno a 20 nucleótidos, como es el caso en los experimentos de miARN-Seq.

La etapa más importante del análisis de datos de RNA-Seq es el alineamiento de las secuencias obtenidas contra un genoma o conjunto de secuencias de referencia. Existen multitud de

programas de software para realizar esta tarea, cada uno con distinta eficiencia computacional (54). Se pueden ajustar múltiples parámetros para aumentar la sensibilidad en detrimento de la velocidad del análisis.

Uno de los parámetros más importantes es la región semilla ("seed"), conceptualmente distinta de la región semilla de la molécula de miARN comentada anteriormente, y que representa el número mínimo de nucleótidos que deben coincidir entre una lectura obtenida y la secuencia de referencia, para que el algoritmo continúe con el proceso de alineamiento tras el emparejamiento con la región semilla. La identificación y posterior extensión de alineamiento a partir de la región semilla es un proceso heurístico, dado que en cada proceso de alineamiento con los mismos datos se van a identificar diferentes regiones semilla de manera aleatoria antes de continuar con la extensión del alineamiento. En RNA-Seq la longitud de la región semilla suele ser de en torno a 20 nucleótidos. Durante la extensión del alineamiento, el algoritmo acepta errores de coincidencia entre nucleótidos ("missmatches"), si bien estos errores implican una penalización que disminuye la puntuación del alineamiento, el cual podría no resultar en un alineamiento válido si el número de errores de coincidencia es muy elevado. Sin embargo, esta penalización es configurable, y por defecto todos los algoritmos aceptan los errores de coincidencia en los alineamientos. La introducción de estos errores es importante para la detección de las variaciones de secuencia y polimorfismos naturalmente presentes en genomas de un mismo organismo. También es importante mencionar las diferencias entre los alineamientos en modo local o global (**Figura 8**). En los alineamientos en modo global, se requiere un alineamiento de principio a fin de la lectura obtenida con las secuencias de referencia, permitiendo la apertura de espacios entre los alineamientos. En los alineamientos en modo local, se permiten alineamientos de sólo una parte de la lectura. La mayoría de programas de software para alineamiento están pre-configurados para aceptar como válidas únicamente las lecturas que cumplen con los requisitos de un alineamiento global.

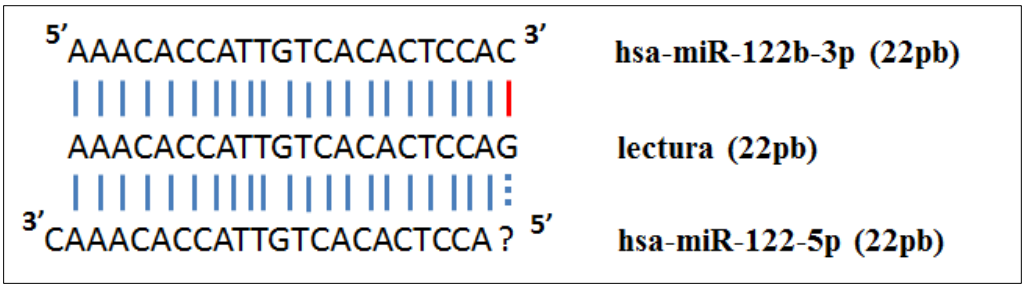


Figura 8: Ilustración de la diferencia entre un alineamiento global y un alineamiento local: la lectura no se alinea de principio a fin con el miARN hsa-miR-122-5p, y aunque las bases anteriores sí lo hacen, no será considerada como una alineación válida en un alineamiento en modo global. Sin embargo la misma lectura si será considerada como un alineamiento válido, ya que alinea de principio a fin con el miARN hsa-miR-122b-3p, siendo considerada la última base como un emparejamiento erróneo, lo cual implica una penalización en el sistema de puntuación del alineamiento.

Tras el proceso de alineamiento de las lecturas obtenidas con las secuencias de referencia, y en base a un sistema de puntuación que depende de la longitud total de la secuencia obtenida y del número de coincidencias con su referencia, se obtiene un fichero con los alineamientos aceptados en formato SAM ("Sequence Alignment Map") o su versión en formato binario BAM (**Figura 9**). El sistema de puntuación se encuentra preconfigurado por defecto en los algoritmos de alineamiento para mostrar en los resultados aceptados únicamente las lecturas que fueron alineadas con mayor fiabilidad. La puntuación de alineamiento se complementa con un valor de calidad del mapeo ("Mapping Quality"), que informa sobre si la lectura ha sido alineada con su referencia de manera única, o bien si fue alineada en distintas ubicaciones.

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUAL	AS	XN	XM	XO	XG	NM	MD	YT	NH	XS
NS500454:244:HHJYGBGX9:4:21505:3147:3806	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:21T0	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:13228:3924	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:19434:4759	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	/AAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:25056:5007	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:23232:5991	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:14692:6128	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:15670:7895	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:21T0	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:18851:8709	0	mmu-let-7c-5p	1	50	18M1I4M	*	0	0	TGAGGT	AAAAA	AS:i:-13	XN:i:0	XM:i:1	XO:i:1	XG:i:1	NM:i:2	MD:Z:19G2	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:12498:8787	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:6897:9140	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:16396:9309	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:16525:9767	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:21T0	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:12365:10346	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+
NS500454:244:HHJYGBGX9:4:21505:4680:10671	0	mmu-let-7c-5p	1	50	22M	*	0	0	TGAGGT	AAAAA	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:22	YT:Z:UU	NH:i:1	XS:A:+

Figura 9: Captura de pantalla mostrando parte de un fichero de lecturas alineadas y aceptadas en formato SAM. Aporta información sobre el nombre de la lectura alineada (QNAME), etiquetas con información sobre la orientación del alineamiento (FLAG), nombre de la secuencia de referencia coincidente (RNAME), posición donde comienza el alineamiento (POS), calidad del mapeo (MPAQ, mejor cuanto mayor es el número), código que informa sobre el número y posición de las bases correctamente emparejadas o desemparejadas (CIGAR), la secuencia de la lectura (SEQ) así como información sobre la calidad de cada base de la lectura en formato ASCII (QUAL).

En el caso de datos de RNA-Seq, al final de todo el proceso de alineamiento, es necesario generar una tabla de conteo, donde se representa el número de lecturas que mapean de la manera más fiel posible a un determinado gen o región del genoma (55). Para el análisis estadístico de estos

datos, generalmente se adopta un modelo específico para datos de distribución discreta, concretamente un modelo de distribución binomial negativa (56), muy similar a la distribución de Poisson pero adaptado a datos con alta dispersión. Es necesario incorporar al modelo una estimación de la dispersión de los valores de conteo obtenidos para cada gen en todas las muestras, procurando reducir al máximo este valor. También es recomendable realizar previamente una etapa de filtrado de los genes que presenten un nivel de expresión residual o nulo en al menos una de las dos condiciones comparadas, ya que al tratarse de un análisis de expresión relativa, conceptualmente no sería correcto evaluar la expresión diferencial de un gen que no se expresa en uno de ambos grupos experimentales.

El proceso de normalización de los datos de RNA-Seq trata de corregir principalmente dos factores técnicos que afectan a la preparación de las muestras secuenciadas. El primero de ellos afecta a la profundidad de cobertura de la secuenciación comentada anteriormente, la cual varía para cada muestra analizada. La librería de análisis utilizada para analizar los datos de RNA-Seq presentados en esta tesis, edgeR (56), realiza esta corrección de manera automática durante el modelado de los datos de conteo para la expresión diferencial. El segundo factor técnico afecta a la composición en ARN de cada muestra analizada, ya que las cantidades analizadas de esta molécula no pueden ser exactamente las mismas en cada celda de la plataforma de secuenciación. Por esta razón el número de lecturas que alinean con un gen determinado no depende sólo del nivel de expresión del gen, sino que también depende la cantidad total de moléculas de ARN presentes en la muestra (57). De este modo, se asume que los genes más expresados en la muestra van a estar representados de manera mayoritaria en la librería a secuenciar, provocando una estimación a la baja del resto de moléculas de ARN presentes en la muestra. La normalización aplicada en este punto utiliza el método TMM (“Trimmed Mean of M-values”) y trata de encontrar un factor de escalamiento para cada muestra secuenciada, con el fin de minimizar las diferencias de conteo de genes entre las muestras que van a analizarse por expresión diferencial. Finalmente, se cuantifica la expresión diferencial de los genes entre 2 condiciones utilizando un test estadístico no paramétrico denominado prueba exacta, y la posterior corrección por testeo simultáneo.

4 MINERÍA DE DATOS NUTRIGENÓMICOS

El reto principal en la era post-genómica actual se centra en organizar, analizar, visualizar e interpretar la gran cantidad de datos que se generan. El uso extendido de las técnicas de análisis de expresión génica de alto rendimiento ha impulsado la aparición de bases de datos públicas para albergar la gran cantidad de información generada. El repositorio de datos más importante dedicado a esta tarea es el americano Gene Expression Omnibus (GEO) (58). Los datos de expresión génica

generados mediante técnicas de alto rendimiento (principalmente arrays de expresión y RNA-Seq) se están acumulando en dichos repositorios públicos de manera exponencial.

Debido al alto coste del análisis de expresión génica a gran escala, los estudios realizados por los laboratorios suelen concentrarse en un tipo celular, tejido, organismo, o compuesto específico, y se realizan utilizando una única plataforma de análisis. Además cuando el investigador publica los datos de un estudio en GEO, los requerimientos de metadatos, es decir los datos que describen los experimentos, son mínimos. Este hecho dificulta la reutilización de los datos de expresión génica en estudios integrativos.

Actualmente los estudios integrativos de experimentos de nutrigenómica en la literatura científica son escasos, así como los trabajos que aplican técnicas de minería de datos con este tipo de estudios (59). En este trabajo se han utilizado dos estrategias para profundizar en el análisis integrativo de los experimentos de nutrigenómica disponibles en GEO: el análisis comparativo de firmas genómicas usando como referencia una base de datos propia de experimentos de nutrigenómica, y la técnica de agrupamiento jerárquico para el descubrimiento de patrones.

4.1 Análisis comparativo de firmas moleculares

4.1.1 Principio del análisis comparativo de firmas moleculares

El fundamento del análisis comparativo de firmas genómicas reside en el supuesto de que un estado biológico se caracteriza por un perfil específico de expresión génica. La aplicación de tratamientos con distintas moléculas en células humanas genera una respuesta transcripcional, caracterizando así los procesos fisiológicos que dichas moléculas activan o reprimen a nivel celular. Una herramienta capaz de comparar patrones de expresión génica debe contener tres elementos principales:

- una colección de perfiles de expresión génica de referencia o base de datos, bien caracterizada en base al tratamiento aplicado para obtener dichos perfiles.
- una interfaz que permita introducir perfiles de expresión génica externos con el fin de compararlos con aquellos incluidos en la base de datos de referencia.
- un algoritmo capaz de comparar el perfil de expresión génica introducido con los presentes en la base de datos y cuantificar la similitud.

La primera aplicación de análisis comparativo de firmas moleculares, utilizando perfiles de expresión génica de manera integrativa para generar nuevo conocimiento, fue propuesta por la plataforma titulada Connectivity Map (CMap) (60). Dicha plataforma alberga en su base de datos firmas moleculares obtenidas tras el tratamiento de distintas células humanas con una gran variedad de medicamentos/drogas, y contiene cientos de perfiles de expresión génica generados a través de una única plataforma de microarray (Affymetrix HT Human Genome U133A).

4.1.2 Conexión entre compuestos y enfermedades

En un principio, el CMap fue diseñado para la tarea de identificar *in silico* nuevos tratamientos terapéuticos para enfermedades. A partir de un perfil de expresión génica característico de enfermedades como la diabetes o el alzhéimer, se pueden identificar drogas capaces de revertir este perfil de expresión, es decir drogas que generan un perfil de expresión opuesto al que caracteriza la enfermedad. De esta manera es posible encontrar relaciones entre las enfermedades y los estados biológicos provocados por la aplicación de distintos tratamientos, de manera que se puedan generar nuevas hipótesis de investigación.

Siguiendo esta estrategia, el CMap ha servido para identificar nuevas indicaciones terapéuticas para compuestos existentes en su base datos, cuya mayoría están aprobados por la Food and Drug Administration (FDA) de EE.UU. Concretamente, gracias al Cmap, un estudio identificó el potencial del compuesto 17-AAG para revertir la firma molecular específica del adenocarcinoma pulmonar. Este compuesto es un potente inhibidor de la proteína de choque térmico HSP90, y de hecho se encuentra en las fases 1 y 2 en varios ensayos clínicos para el tratamiento de varios cánceres (61). Siguiendo la misma estrategia, otro estudio consiguió identificar la rapamicina como un compuesto potencial para revertir la resistencia a glucocorticoides en tratamientos contra el cáncer, sugiriendo así su uso combinado en estos tratamientos (62). En estos supuestos, el análisis comparativo de firmas moleculares se basa en la correlación negativa entre dos perfiles de expresión génica.

4.1.3 Conexión entre los modos de acción de los compuestos

La estrategia de comparación de firmas moleculares también presenta un enorme potencial para elucidar el modo de acción de los compuestos que causan determinados fenotipos. Es posible identificar drogas que comparten un mismo modo de acción si el perfil de expresión génica que provocan es suficientemente parecido. De esta manera es posible identificar mecanismos moleculares y vías metabólicas en común, activados o reprimidos por diferentes compuestos.

Esta metodología presenta un gran potencial para el reposicionamiento de drogas/medicamentos existentes. Del mismo modo, dado que la mayoría de drogas/medicamentos no suelen actuar únicamente en una diana molecular, se puede usar este tipo de análisis para encontrar moléculas que presentan menores efectos adversos a la hora de tratar ciertas enfermedades (63). Con el objetivo de ahondar en los mecanismos moleculares responsables de los modos de acción, es importante incorporar información sobre vías metabólicas o funciones moleculares que corresponden a los genes que dos compuestos distintos activan o reprimen de manera común. En estos supuestos, el análisis comparativo entre firmas moleculares se basa en su correlación positiva.

4.1.4 Aplicaciones web existentes y limitaciones

Existen herramientas que intentan extraer automáticamente firmas moleculares a partir de los datos presentes en GEO (64,65), pero debido a la inconsistencia de la anotación de los estudios presentes en los repositorios públicos, estas tienden a cometer errores durante la asignación automática de las muestras a los grupos control y tratamiento. Es por ello que la obtención de firmas moleculares de manera manual a partir de los experimentos disponibles en GEO presenta múltiples ventajas para generar nuevas hipótesis de investigación fiables (66).

También existen aplicaciones de la metodología de comparación de firmas moleculares enfocadas en la investigación de enfermedades raras, dado que la financiación para investigar este tipo de enfermedades es escasa. Se basan en un principio similar al de los orígenes del CMap, asumiendo que el perfil de expresión génica propio de una enfermedad rara puede ser revertido por drogas cuyo tratamiento provoque un perfil de expresión opuesto (67). Sin embargo los autores admiten que este enfoque puede resultar ingenuo para el reposicionamiento de medicamentos dado la complejidad de los sistemas biológicos de alto nivel.

4.1.5 ALGORITMOS PARA COMPARACIÓN DE FIRMAS MOLECULARES

El algoritmo de análisis comparativo de firmas moleculares implementado en las herramientas citadas anteriormente proviene originalmente del método computacional de análisis de enriquecimiento de conjuntos de genes (GSEA) (68). A partir de dos listas de genes, ordenados por su nivel de expresión diferencial, este algoritmo primero encuentra los genes coincidentes entre ambas listas, y después evalúa si estas coincidencias se reparten de manera aleatoria a lo largo de ambas listas, o bien si las coincidencias se sitúan más bien en los extremos de ambas listas (genes mayormente sobreexpresados o reprimidos) (**Figura 10**). Para ello se calcula una puntuación de enriquecimiento (ES). Un valor de ES positivo significa que se detecta un enriquecimiento de genes en común entre ambas listas en el extremo superior (sobreexpresados) y viceversa para un ES negativo. Un valor de ES cercano a cero representa un enriquecimiento aleatorio.

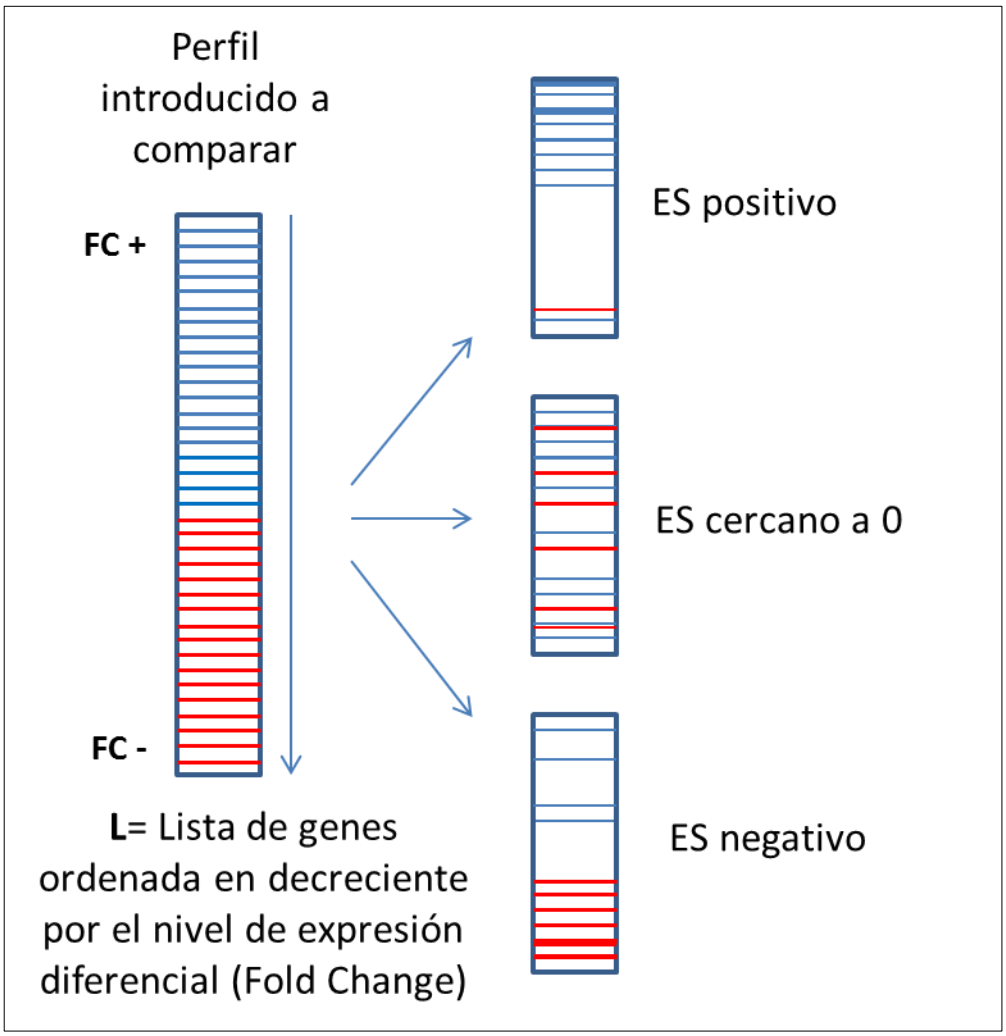


Figura 10: Ilustración de los tres posibles casos de puntuación de enriquecimiento entre una lista de genes L (firma molecular a comparar) y los perfiles de referencia. Los colores azul y rojo representan genes sobreexpresados y reprimidos, respectivamente. La obtención de un perfil de expresión de referencia con ES positivo indica una alta coincidencia de genes mayormente sobreexpresados entre el perfil de referencia y el perfil de expresión introducido, y viceversa para un ES negativo.

El cálculo del ES parte de la lista de genes que se quiere comparar (lista L), ordenada de manera descendente por nivel de expresión. Se empieza recorriendo la lista desde su principio (genes altamente sobreexpresados), y el valor de la puntuación incrementa cuando se encuentra un gen en común entre la lista ordenada y un perfil de expresión en la referencia (**Figura 11**). Este incremento es positivo si la coincidencia aparece en el extremo superior de ambas listas, y viceversa. El incremento de la puntuación se realiza de manera ponderada, ya que los genes coincidentes que se encuentran en los extremos de ambas listas obtienen mayor puntuación. Paralelamente, cuando se evalúan los genes sobreexpresados, la puntuación ES decrece progresivamente cuando uno de los genes de la lista ordenada no se encuentra en la lista de referencia, y viceversa para los genes reprimidos. Tras finalizar este proceso, la desviación máxima del valor 0 representa el ES entre dos firmas moleculares.

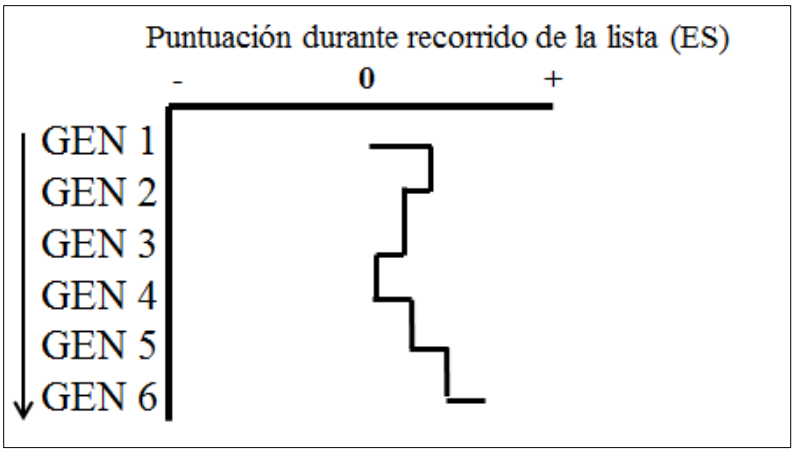


Figura 11: Esquema del proceso de cálculo del ES, empezando desde el extremo superior de la lista de genes (sobreexpresados). Los genes 1,4,5 y 6 de la lista L coinciden con la firma molecular de referencia. El gen 1 aporta el mayor incremento de ES ya que la coincidencia aparece en las primeras posiciones de ambas listas de genes (altamente sobreexpresado). Los genes 2 y 3 no coinciden entre ambas listas, con lo cual el valor ES decrece ligeramente.

En los resultados es recomendable representar el valor de la puntuación de enriquecimiento normalizada (NES), el cual tiene en cuenta el tamaño de los perfiles de referencia incluidos en la base de datos, así como la correlación entre estos y el perfil de expresión interrogado (lista L). Para el cálculo, se divide la puntuación ES obtenida para cada perfil de referencia, por la media de las ES calculadas para otros perfiles de referencia permutados, de mismo tamaño y símbolo de ES que el perfil de referencia a normalizar. La fórmula es la siguiente:

NES = ES / media (ES contra permutaciones de Perfiles de referencia)

En este caso, dónde se trabaja directamente con una lista L ordenada, la estimación de la significancia estadística de los NES obtenidos para cada perfil de referencia se calcula creando

perfiles de referencia aleatorios a partir de los presentes en la base datos, y de mismo tamaño que el perfil de referencia evaluado. De esta manera se obtienen nuevos valores de NES, generando una distribución nula de estos valores, y luego se calcula un valor *p* evaluando el número de veces que el nuevo valor NES obtenido ha sido mejor que el original.

4.1.6 Base de datos de firmas moleculares de referencia

Otro aspecto fundamental en el análisis comparativo de firmas moleculares es la correcta selección de los genes que componen las firmas de referencia, y que caracterizan el estado biológico provocado tras un tratamiento determinado. A pesar del gran impacto del CMap, las firmas moleculares incluidas en su base de datos cuentan con un bajo nivel de replicación, ya que generalmente cada experimento incluye una única muestra de tratamiento y control, con lo cual el orden de los genes analizados en las listas obtenidas, que representan las firmas moleculares, no está basado en un valor de significancia estadística, sino únicamente en su nivel de expresión diferencial. Una selección errónea de los genes que componen las firmas moleculares de referencia puede llevar a identificar similitudes entre el diseño experimental de dos experimentos, en lugar de conexiones con sentido biológico. Esto ha sido discutido en la literatura científica y se han propuesto métodos más robustos para la construcción de las firmas moleculares de referencia, así como para la evaluación de la significancia estadística de las conexiones encontradas (69).

Por el contrario, en la iniciativa GSEA, la base de datos de referencia está compuesta por listas de genes ordenadas resultantes de experimentos de expresión diferencial con alta replicación, los conjuntos de genes ("GeneSets"). Estos conjuntos de genes son el resultado de experimentos concretos, y tras el proceso de filtrado en base a criterios estadísticos, representan únicamente los genes con mayor nivel de expresión diferencial y mayor significancia estadística. Actualmente esta base de datos cuenta con 22.596 conjuntos de genes (octubre 2019), de tamaño variable, divididos por temáticas funcionales (vías metabólicas, ontologías funcionales, firmas oncogénicas etc.)

En la actualidad no existe ninguna base de datos como la que se ha generado durante este trabajo de doctorado, y que recopile conjuntos de genes diferencialmente expresados tras tratamientos celulares utilizando alimentos y sus compuestos bioactivos.

4.2 Agrupamiento jerárquico

Las técnicas de agrupamiento ("clustering") son un tipo de métodos dentro del área del aprendizaje automático no supervisado. A partir de un conjunto de datos de entrada, estas técnicas tienen el objetivo de agrupar los datos en base a su similitud o disimilitud, sin tener en cuenta

información previa sobre el grupo o la clase a la cual pertenecen los datos del conjunto. El resultado es una representación gráfica de los resultados de agrupamiento en forma de dendrograma.

En el campo de la biología molecular, para los datos obtenidos por experimentos de transcriptómica, la representación gráfica más extendida para el descubrimiento de patrones es el mapa de calor agrupado ("clustered heatmap"). Esta visualización ha demostrado su amplio potencial para el descubrimiento de patrones de expresión. Por ejemplo se ha demostrado que el mapa de calor agrupado permite clasificar correctamente tumores humanos en base al perfil de expresión de los miARN's (70). El resultado es un agrupamiento jerárquico de los datos en dos dimensiones, donde generalmente las columnas representan las muestras/experimentos y las filas representan los genes analizados. La relación entre los grupos en ambas dimensiones se representa mediante un dendrograma, adyacente a cada dimensión, y que define las distancias entre los grupos. Los colores del mapa de calor indican correlaciones entre los genes o las muestras incluidas en el análisis, y pueden revelar potenciales relaciones funcionales entre los mismos.

4.2.1 Algoritmo de agrupamiento jerárquico

Los algoritmos para el agrupamiento jerárquico de datos de transcriptómica se basan en tres etapas esenciales:

- cálculo de las matrices de distancias entre genes o muestras.
- estrategia de agrupamiento: aglomerativo o divisivo.
- método de vinculación entre grupos.

La matriz de distancias informa sobre la similitud entre todos los pares de vectores que corresponden a los genes/muestras que se quieren incluir en la visualización de mapa de calor agrupado. Es importante definir qué métrica se va a utilizar para evaluar estas distancias. Los métodos de cálculo de distancia utilizan distintas fórmulas matemáticas y se usan uno u otro método dependiendo de los datos a analizar, considerando la homogeneidad de la escala para todo el conjunto de datos y también la presencia de valores extremos. Para experimentos de transcriptómica, el método usado comúnmente es el de la distancia euclídea (**Figura 12**). La fórmula matemática para el cálculo de la distancia euclídea entre 2 vectores de datos p y q es la siguiente:

$$\mathrm{d}(\mathrm{p},\mathrm{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

	BIRC5	CENPF	CKS1B	KIF2C	MAD2L1	MCM3	NUP107	POLD4	RFC4
BIRC5	0.000000	2.143463	1.992819	1.4221320	1.4464263	2.0372113	2.0653043	6.237357	1.7829155
CENPF	2.143463	0.000000	3.378998	1.0030108	1.6238086	2.0285852	3.0585226	7.500080	2.6419999
CKS1B	1.992819	3.378998	0.000000	2.5657498	2.0537609	2.2722242	1.3915723	4.832973	1.3029292
KIF2C	1.422132	1.003011	2.565750	0.0000000	0.8441865	1.4509717	2.3360096	6.798563	1.8053614
MAD2L1	1.446426	1.623809	2.053761	0.8441865	0.0000000	1.1977937	1.8589621	6.487537	1.3580074
MCM3	2.037211	2.028585	2.272224	1.4509717	1.1977937	0.0000000	1.4422005	6.305940	1.2317701
NUP107	2.065304	3.058523	1.391572	2.3360096	1.8589621	1.4422005	0.0000000	5.181420	0.9820027
POLD4	6.237357	7.500080	4.832973	6.7985631	6.4875366	6.3059401	5.1814205	0.000000	5.3776040
RFC4	1.782915	2.642000	1.302929	1.8053614	1.3580074	1.2317701	0.9820027	5.377604	0.0000000

Figura 12: Ejemplo de una matriz de distancias obtenida por el cálculo de la distancia euclídea entre los valores de log2 FC de 9 genes. 0 representa la similitud absoluta, y mayor es la similitud entre dos pares cuando menor es el valor de distancia.

Una vez obtenidas las distancias entre los vectores de valores para cada gen/muestra, se define si el tipo de agrupamiento se va a realizar de manera aglomerativa o divisiva. En el agrupamiento aglomerativo, cada vector de datos correspondiente a los genes/muestras es asignado a un clúster individual, y estos se van agrupando progresivamente en función de su similitud en base a la matriz de distancia obtenida previamente. En cambio en el agrupamiento divisivo, todos los vectores de datos de genes/muestras se asignan a un mismo clúster, y este se va particionando en grupos con mayor disimilitud, hasta alcanzar el mismo número de grupos que de genes/muestras hay en los datos (**Figura 13**). El agrupamiento aglomerativo es el más usado para los mapas de calor en experimentos de transcriptómica, siendo muy útil para identificar clústeres pequeños, contrariamente al agrupamiento divisivo, el cual es útil para identificar clústeres más amplios.

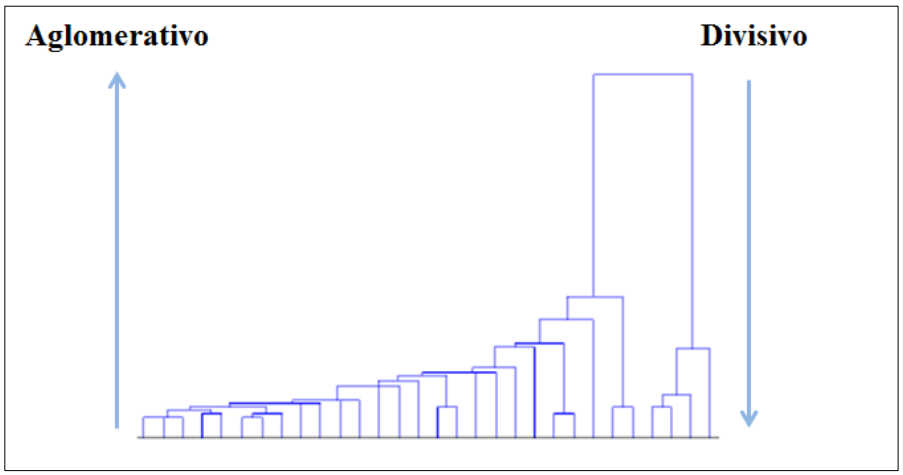


Figura 13: Ilustración de los dos tipos de agrupamiento jerárquico posibles

La etapa final trata sobre la manera en la que los clústeres individuales y los nuevamente formados se van a ir enlazando entre ellos de manera aglomerativa en base a la matriz de distancias, la cual se irá actualizando con la distancia correspondiente a los nuevos clústeres que se han ido formando. Los métodos más comunes para esta tarea son el enlazamiento completo, individual, medio y centroide (**Figura 14**). Los pares de clústeres que presenten la menor distancia en cualquiera de estos cuatro métodos serán enlazados, resultando en nuevas ramas del dendrograma.

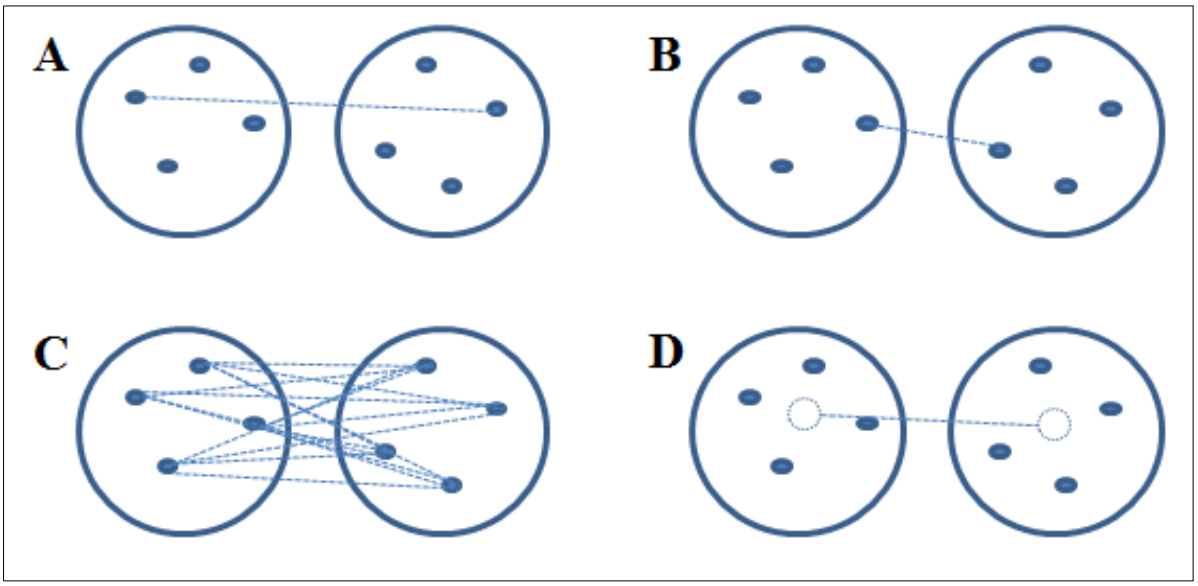


Figura 14: Ilustración de los métodos utilizados para enlazar clústeres. A - Completo, se basa en la distancia mayor entre 2 valores de ambos clústeres. B - Individual, se basa en la distancia menor entre 2 valores de ambos clústeres. C - Media, se basa en la distancia media entre todos los valores de ambos clústeres. D - Tras encontrar el centroide, que representa la proyección de la media aritmética de todos los valores del clúster, se utiliza la distancia entre ambos centroides.

4.2.2 Mapa de calor

Además de la información sobre la similitud entre genes/muestras que proporcionan los dendrogramas obtenidos, los valores individuales de la matriz de datos utilizada para el agrupamiento pueden representarse en base a una escala de colores. De esta manera es posible identificar correlaciones y patrones entre genes/muestras de manera visual. La intensidad de los colores visualizados correlaciona directamente con el nivel de expresión de los genes analizados, o su expresión diferencial en un experimento, con lo cual el conocimiento de la escala en la cual los datos se representan es fundamental. Por esta razón es común utilizar datos transformados a escala logarítmica con el objetivo de reducir la variabilidad de la matriz de datos. Del mismo modo, resulta muy útil tipificar los datos contenidos en los vectores de la matriz del mapa de calor, restando la

media y dividiendo por la desviación típica cada valor presente en el vector de datos (valor *z*). Estas transformaciones matemáticas no alteran los perfiles de variación presentes en la matriz de datos original, y facilita la identificación visual de patrones en los datos (**Figura 15**).

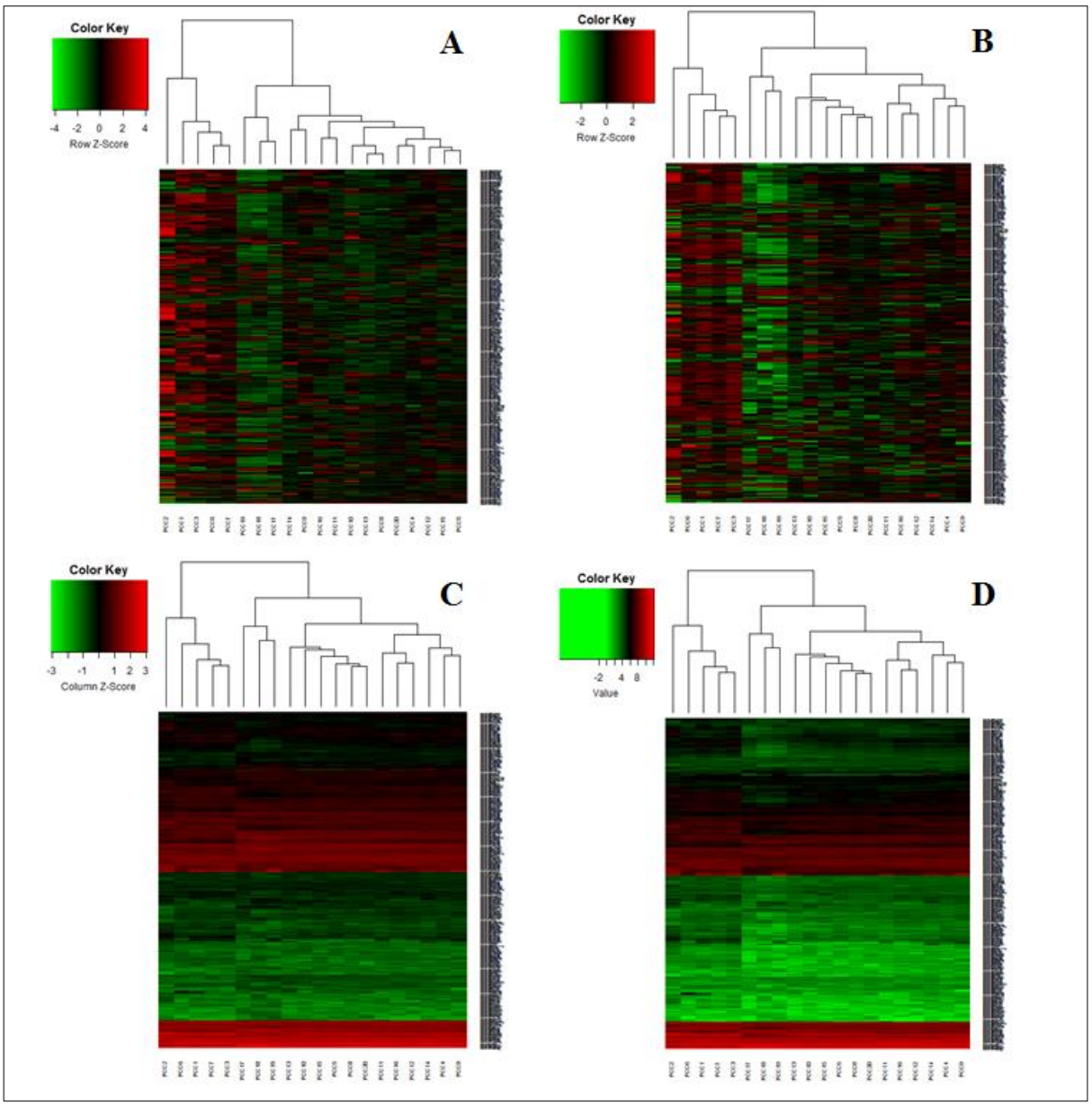


Figura 15: Ejemplo de la influencia de la escala y tipificación de los datos para la identificación de patrones en los mapas de calor. A - Valores tipificados por filas y en escala lineal. B- Valores en tipificados por filas y en escala logarítmica. C- Valores tipificados por columnas y en escala logarítmica. D - Valores absolutos sin tratamiento. La transformación a escala logarítmica y tipificación por filas (B) mejora la identificación de correlaciones entre genes mediante inspección visual.

Para una correcta interpretación del mapa de calor, es importante conocer la escala de los datos visualizados, así como indicar qué valores se están visualizando en el mapa de calor. Para experimentos de transcriptómica, generalmente se visualizan los valores de expresión normalizados

de las muestras incluidas. Del mismo modo se pueden visualizar valores del cambio proporcional en logaritmo en base 2 (log2 FC) del resultado de expresión diferencial tras el análisis de experimentos.

Una limitación del mapa de calor en transcriptómica es que, si bien permite identificar patrones a nivel global, las relaciones entre subgrupos de genes/muestras pasan desapercibidas. Para relaciones entre subgrupos, la técnica de biclustering es ideal y ha demostrado su potencial para la identificación de subtipos de tumores (71).

HIPÓTESIS Y OBJETIVOS

La hipótesis de esta tesis doctoral es que las propiedades saludables de determinados alimentos o productos alimentarios pueden explicarse por la capacidad de algunos de sus componentes para regular diversas rutas metabólicas al afectar a la expresión de genes. En un mayor grado de concreción, la hipótesis considera que los mecanismos moleculares de los efectos saludables de los componentes de los alimentos pueden ser frecuentemente coincidentes con los de drogas empleadas por la Medicina Personalizada.

En consecuencia, el objetivo general de esta tesis doctoral es recopilar, analizar e integrar datos nutrigenómicos experimentales con el fin de diseñar una herramienta bioinformática para la identificación de las bases moleculares de los efectos saludables de los alimentos o productos alimentarios.

Los objetivos específicos de esta tesis son tres:

1) Analizar datos resultantes de experimentos de nutrigenómica utilizando las técnicas de transcriptómica más actuales, con el fin de elucidar los efectos del consumo de una dieta suplementada con hidroxitirosol en la expresión de miARN's en el hígado de ratón.

2) Llevar a cabo la creación y análisis integrativo de una base de datos de firmas moleculares generada a partir de datos de experimentos de nutrigenómica, realizados en células humanas y disponibles públicamente en el repositorio de datos GEO.

3) Desarrollar una aplicación web, apoyada en la nueva base de datos creada, para aplicar técnicas de minería de datos con el fin de elucidar los mecanismos moleculares que confieren las propiedades saludables a determinados alimentos y sus compuestos bioactivos.

En base a los tres objetivos específicos, esta tesis se divide en 3 capítulos:

Capítulo 1: Análisis experimental del efecto del hidroxitirosol en la expresión de micro ARN's en el hígado de ratón (Objetivo 1).

Capítulo 2: Creación y análisis integrativo de una base de datos a partir de experimentos de nutrigenómica en células humanas (Objetivo 2).

Capítulo 3: Desarrollo de una aplicación web para minería de datos en nutrigenómica (Objetivo 3)

MATERIALES Y MÉTODOS

CAPÍTULO 1: ANÁLISIS EXPERIMENTAL DEL EFECTO DEL HIDROXITIROSOL EN LA EXPRESIÓN DE MICRO ARN'S EN EL HÍGADO DE RATÓN.

1. MATERIALES

El compuesto HT fue donado por Seprox Biotech (Madrid, España). Las dietas control purificadas fueron obtenidas de Research Diets Inc., New Brunswick, NJ (Estados Unidos de América).

2. ANIMALES Y DIETAS

En este experimento de nutrigenómica se han utilizado 14 ratones de la cepa C57BL/6 de 2 meses de edad. Una semana antes del comienzo del experimento, los ratones fueron aclimatados a un ciclo de luz/sombra de 12h, siendo el ciclo de luz entre las 7a.m. y las 7p.m., y consumiendo una dieta standard "chow" y agua *ad libitum*. Posteriormente, y durante 8 semanas, 7 ratones fueron alimentados con una dieta control purificada, y otros 7 con la misma dieta suplementada en HT. La dosis de suplementación de HT fue de 45 mg/kg, la cual se aproxima a la dosis que un humano podría consumir a diario. Cada dieta control purificada aporta respectivamente 24.0%, 15.0% y 61.0% del total de kilocalorías de proteínas, grasas y carbohidratos.

Los animales fueron sacrificados entre las 10a.m. y las 11a.m. con el objetivo de reducir las posibles variaciones diurnas intrasujeto. Los tejidos de hígado fueron extraídos, lavados en tampón fosfato salino, congelados en nitrógeno líquido y guardados a -80ºC.

3. CONSIDERACIONES ÉTICAS

Este estudio con animales fue aprobado por los comités de ética de la Universidad Complutense de Madrid (CEA-UCM 93/2012), de la Universidad de Lleida (CEEA 10-06/14) y de la Universidad Mixta de investigación de Zaragoza. Todos los procedimientos han cumplido con las direcciones de la Guía para el Cuidado y Uso de animales de laboratorio.

4. EXTRACCIÓN DE ARN Y SECUENCIACIÓN DE MICRO ARN'S

El ARN total de cada tejido fue extraído mediante el uso del reactivo de lisis Qiazol y las columnas Mini kit miRNeasy® (Qiagen, Madrid, España). Posteriormente fue cuantificado utilizando el espectrofotómetro NanoDrop-1000 (Thermo Fisher Scientific Inc., España). La integridad del ARN extraído se comprobó mediante el bioanalizador Agilent 2100, y las muestras con un valor de RIN >8 fueron seleccionadas para su análisis posterior.

Para la obtención de ADNc a partir del ARN total extraído se ha utilizado el kit Universal cDNA synthesis kit II (Exiqon). Para la preparación de las librerías a secuenciar, se ha utilizado el kit

NEBNext® 204 multiplex small RNA Library Prep Set (New England BioLabs, Ipswich, MA, Estados Unidos de América). La secuenciación se ha realizado en la plataforma NextSeq 500 (Illumina).

5. ANÁLISIS DE LOS DATOS DE SECUENCIACIÓN DE MICRO ARN'S

En primer lugar, se eliminaron de las lecturas obtenidas las secuencias correspondientes a los adaptadores de secuenciación propios de Illumina. Tras ello se ha usado el software Bowtie 2 (72) para alinear las lecturas de secuenciación contra un conjunto de 816 secuencias de referencia de alta confidencia, correspondientes a miARN's maduros de ratón, y obtenidas de la base de datos miRBase v21.1 (37). Para la cuantificación de los niveles de expresión de los miARN's detectados, se han considerado las lecturas alineadas de manera única con las secuencias de referencia.

6. ANÁLISIS FUNCIONAL DE LOS MICRO ARN'S

A partir de los miARN's diferencialmente expresados, se han obtenido sus posibles dianas moleculares. Para ello se ha utilizado la aplicación miRWalk 3.0 (73), limitando los resultados a las posibles dianas con interacción en la posición 3'UTR, y con un valor del parámetro "Binding P-value" igual a 1. A partir de la lista de genes identificados, y que son diana de al menos dos miARN's diferencialmente expresados, se ha realizado un enriquecimiento funcional de las vías metabólicas que se encuentran sobre representadas utilizando la base de datos de Panther (74). Se ha utilizado el test exacto de Fisher para evaluar la significancia estadística de los resultados, y los valores *p* obtenidos han sido corregidos por el método de Benjamini-Hochberg (FDR) (52), conservando únicamente los resultados con un valor FDR<=0.05. La representación gráfica de la red de interacciones miARN's – ARNm's identificada se ha realizado utilizando el software Cytoscape versión 3.4., y contiene únicamente los genes que son diana de al menos dos miARN's diferencialmente expresados.

CAPÍTULO 2: CREACIÓN Y ANÁLISIS INTEGRATIVO DE UNA BASE DE DATOS A PARTIR DE EXPERIMENTOS DE NUTRIGENÓMICA EN CÉLULAS HUMANAS.

1. RECOPILACIÓN DE DATOS

Para la creación de una base de datos de experimentos de nutrigenómica, se han recopilado datos de expresión correspondientes a ensayos realizados en células humanas, analizados utilizando arrays de expresión, y con un diseño experimental que incluya al menos dos réplicas por cada grupo experimental. Utilizando la base de datos GEO, se han realizado consultas utilizando los siguientes términos de búsqueda para la identificación de estudios de nutrigenómica: "nutrigenomics", "nutrient", "nutrition", "extract", "natural product" y "phytochemical". Los estudios identificados están compuestos por uno o múltiples experimentos.

2. TRATAMIENTO DE LOS DATOS DE EXPRESIÓN IDENTIFICADOS

Para cada estudio identificado, tras inspeccionar cuidadosamente el correspondiente diseño experimental, las muestras han sido asignadas manualmente a los grupos tratamiento o control. Para datos de expresión generados con la plataforma Affymetrix, se han descargado los datos crudos (en formato CEL) para ser normalizados de manera local, utilizando el software R y el algoritmo RMA (49), disponible en la librería "affy" de Bioconductor. En cuanto a los datos generados con el resto de arrays de expresión, se han descargado directamente las matrices de datos de expresión normalizadas y disponibles en GEO.

Posteriormente se han anotado las sondas de cada experimento analizado con su correspondiente nombre de gen, utilizando el software R y la librería "annotate". Para el análisis de expresión diferencial de cada experimento identificado, se ha utilizado el software R y la librería "limma" (50). En el caso de los genes duplicados obtenidos en la matriz de expresión diferencial, resultantes de distintos "probesets" que analizan la expresión de un mismo gen, se han obtenido las medias de los correspondientes valores de expresión normalizados.

3. CONSTRUCCIÓN DE LA BASE DE DATOS

La primera versión de la base de datos, generada con un conjunto de experimentos de nutrigenómica recopilados hasta Enero de 2017, se ha construido a partir de los niveles de expresión diferencial de todos los genes analizados en los arrays de expresión: los valores log2 FC. Esta versión de la base de datos contiene los resultados de un total de 81 experimentos. Dado que los datos incluidos habían sido generados en 19 tipos de plataformas de arrays de expresión distintas, únicamente se han incluido los datos de expresión de los genes analizados en todas las plataformas, obteniendo así una matriz de datos sin valores perdidos. De este modo, ha sido necesario eliminar

experimentos realizados con plataformas que generaban un mayor número de datos perdidos en relación a la cantidad total de genes analizados, obteniendo finalmente una matriz de datos con 4.649 genes en las filas, y 73 experimentos. Los estudios de origen y los detalles experimentales se encuentran en la **tabla 3**. A partir de estos datos, se han realizado los análisis integrativos de agrupamiento jerárquico.

4. AGRUPAMIENTO JERÁRQUICO Y MAPAS DE CALOR

Para los gráficos de agrupamiento jerárquico y mapas de calor generados durante el análisis integrativo de los datos de expresión diferencial, se ha utilizado el software R y la librería "heatmap.2". A partir de los valores de expresión diferencial, primero se ha calculado la distancia euclídea para definir la relación entre los experimentos. El método elegido para la construcción de los clústeres ha sido de tipo aglomerativo y completo.

5. ANÁLISIS FUNCIONAL DE GENES

El análisis de enriquecimiento funcional de los genes identificados durante el análisis integrativo ha sido realizado utilizando la aplicación Genecodis 3 (75) y anotaciones ontológicas de procesos biológicos. Se ha empleado un test hipergeométrico para evaluar la significancia estadística de los resultados, y los valores *p* han sido corregidos por el método de Benjamini-Hochberg (FDR) (52), conservando únicamente los resultados con un valor FDR ≤ 0.05.

Tabla 3: Detalles sobre los experimentos de nutrigenómica incluidos en la base de datos (Enero 2017)

ID GEO	Experimentos	Alimentos y compuestos bioactivos	Líneas celulares	Plataforma
GSE14204	1	25-hydroxycholesterol	HUH7	GPL96
GSE15322	1	Extracto de naranja enriquecido en isoflavonas	CCD-18Co	GPL570
GSE18589	4	Ácidos oleico, palmítico, eicosapentaenoico y linoleico	Miotubos	GPL7020
GSE18741	3	Lactobacillus acidophilus, L. casei and L. rhamnosus	Duodeno	GPL6791
GSE20940	4	Lactobacillus rhamnosus	Macrófagos	GPL6879
GSE21976	2	Bifidobacterium bifidum	Caco2	GPL7020
GSE23630	3	Lactobacillus casei	Explantes intestinales	GPL570
GSE25412	2	Resveratrol	MCF7	GPL6244
GSE28384	4	Amorfrutinas	Adipocitos	GPL6947
GSE28813	1	Sulforafano	MCF10A	GPL4133
GSE34175	3	Lactobacillus acidophilus	Caco2	GPL6884
GSE38227	1	Extracto de nuez de areca	Queratinocitos	GPL4133
GSE39828	1	Embelina	HEK 293T	GPL6480
GSE43166	4	Folato y vitamina B12	CHUB-S7	GPL14550
GSE44290	3	Extracto de hojas de Achyranthes aspera	PaCa-2	GPL4133
GSE45357	1	Gliadina	Caco2	GPL4133
GSE45804	1	Ácido lacciaco A	MCF7	GPL6244
GSE48668	3	Tocotrienoles	HeLa	GPL7020
GSE50945	2	Aceites de pescado y soja	Células de placenta	GPL10558
GSE53049	2	Witaferina A	MCF7, MDA	GPL10558
GSE55897	6	Indol-3-carbinol	MCF7, MDA, T47D, ZR75	GPL10904
GSE56265	3	Ácido lisofosfatídico	MCF7. MDA, PC3	GPL570
GSE56496	3	Extracto de romero	SW620	GPL6480
GSE58749	8	Ácido retinoico y ácido 3,4-didehidroretinoico	Queratinocitos	GPL17692
GSE65397	2	Ácidos grados de yema de huevo	MCF7	GPL17077
GSE65527	1	Urolitina A	LNCaP	GPL13667
GSE65722	2	Extractos de romero y ácido carnósico	HT29	GPL16686
GSE7259	1	Quercetina	Caco2	GPL571
GSE73650	1	Transresveratrol	Caco2	GPL16686

CAPÍTULO 3: DESARROLLO DE UNA APLICACIÓN WEB PARA MINERÍA DE DATOS EN NUTRIGENÓMICA.

1. ACTUALIZACIÓN DE LA BASE DE DATOS DE EXPERIMENTOS DE NUTRIGENÓMICA

La segunda versión de la base datos de experimentos de nutrigenómica ha sido generada a partir de la primera versión (**Capítulo 2**), añadiendo experimentos de nutrigenómica que se han seguido recopilando hasta Marzo de 2018 (**Figura 16**). Los estudios incluidos tras la actualización, así como los correspondientes detalles experimentales, se encuentran disponibles en la **tabla 4**. El tratamiento de los nuevos datos ha sido idéntico al realizado para la construcción de la primera base de datos (**Capítulo 2**). Con el objetivo de definir cada experimento incluido mediante un conjunto de genes diferencialmente expresados, tras el análisis de expresión diferencial se han seleccionado el 10% del total de genes incluidos en el array de expresión utilizado, ordenados por su nivel de significancia estadística (columna de valor *p* corregido) y sin ninguna etapa de filtrado previa. De esta manera se ha obtenido una matriz de datos de valores log2 FC, con 41.148 filas (incluyendo genes, locis sin nombre de gen y ARN's no codificantes) y 231 experimentos de nutrigenómica (columnas). De este modo cada experimento se encuentra definido por una firma molecular o conjunto de genes ("Gene Sets"), caracterizada por el nivel de expresión diferencial del 10% del total de genes analizados en el array y con mayor significancia estadística. Esta base de datos se ha utilizado para construir los 2 módulos de la aplicación web NutriGenomeDB.

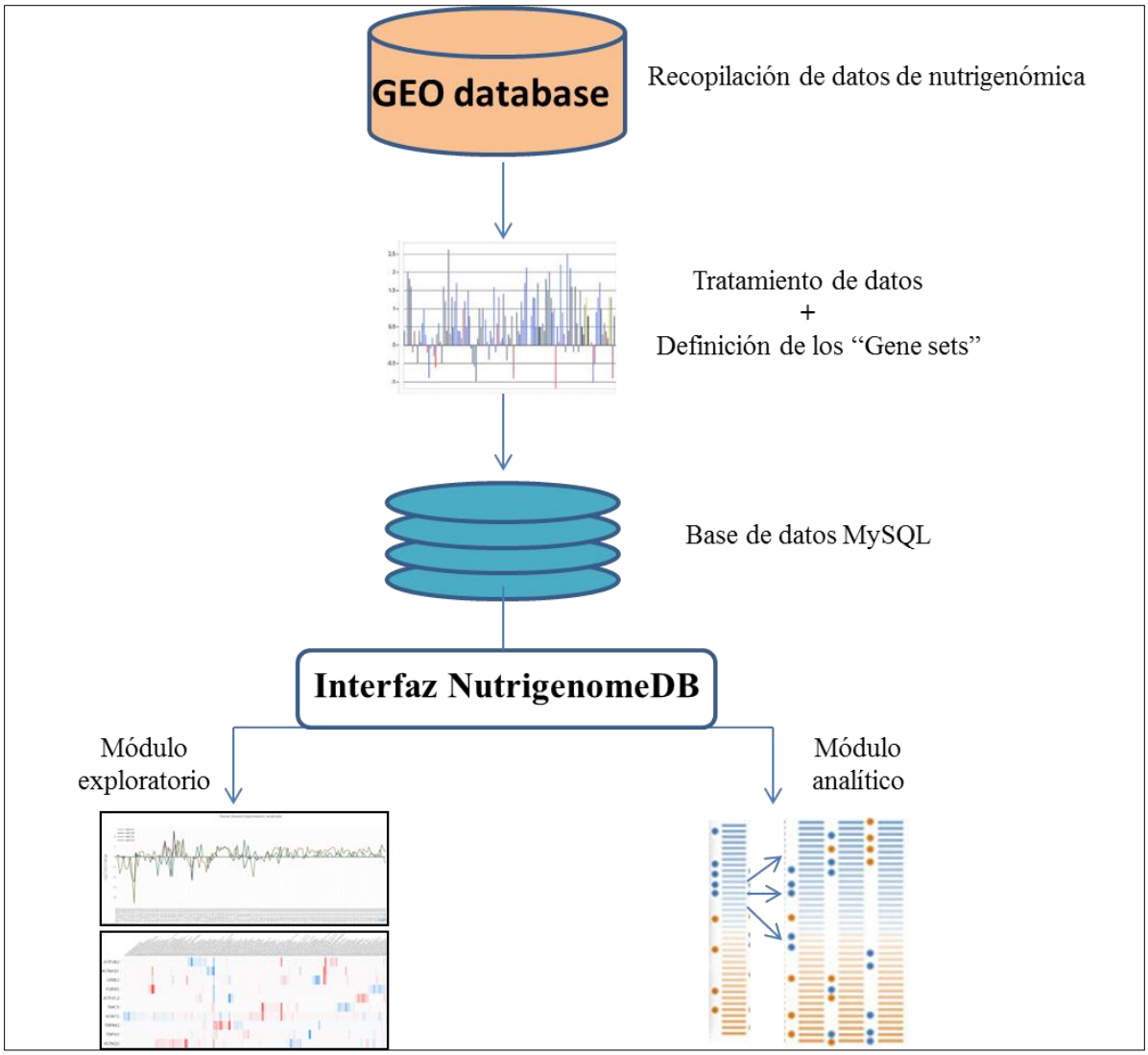


Figura 16: Visión general del flujo de trabajo empleado para desarrollar la aplicación web para minería de datos en nutrigenómica. El término "Gene Sets" se refiere a las firmas moleculares de referencia incluidas en la base de datos.

Tabla 4: Detalles sobre los experimentos de nutrigenómica incluidos en la base de datos (Marzo 2018)

ID GEO	Experimentos	Alimentos y compuestos bioactivos	Líneas celulares	Plataforma
GSE100224	1	Extracto de Wedelia chinensis	22Rv1	GPL10558
GSE100846	2	Extractos de mora y de Rubus vagabundus	SK-N-MC	GPL15034
GSE102891	2	Extracto de Boswellia serrata y ácido 3-O-Acetyl-B-boswellico	MDA-MB-231	GPL17692
GSE104268	2	Tricostatina A, extracto de semilla de uva	SK-MEL-3	GPL17692
GSE10896	2	Curcumina	U937	GPL570
GSE109607	10	Extracto de semilla de uva, proantocianidinas oligoméricas	SW620, SW480, RKO, HT29 and HCT116	GPL16791
GSE14204	1	25-hydroxycholesterol	HUH7	GPL96
GSE15322	1	Extracto de naranja enriquecido en isoflavonas	CCD-18Co	GPL570
GSE18589	4	Ácidos oleico, palmítico, eicosapentaenoico y linoleico	Miotubos	GPL7020

ID GEO	Experimentos	Alimentos y compuestos bioactivos	Líneas celulares	Plataforma
GSE18741	3	Lactobacillus acidophilus, L. casei and L. rhamnosus	Duodeno	GPL6791
GSE20114	1	Ácido docosahexaenoico	Glóbulos blancos	GPL570
GSE20940	4	Lactobacillus rhamnosus	Macrófagos	GPL6879
GSE21976	2	Bifidobacterium bifidum	Caco2	GPL7020
GSE23630	3	Lactobacillus casei	Explantes intestinales	GPL570
GSE25412	2	Resveratrol	MCF7	GPL6244
GSE28384	4	Amorfrutinas	Adipocitos	GPL6947
GSE28813	1	Sulforafano	MCF10A	GPL4133
GSE32357	1	Resveratrol	Células musculares	GPL11532
GSE34074	9	Genisteina	Fibroblastos	GPL6947
GSE34175	3	Lactobacillus acidophilus	Caco2	GPL6884
GSE35382	2	Café y ácido caféico	HT29	GPL570
GSE37089	2	Ácido retinoico	Células epiteliales	GPL10558
GSE38227	1	Extracto de nuez de areca	Queratinocitos	GPL4133
GSE38833	2	Hidroxitirosol y etil éter de hidroxitirosil	Caco2	GPL570
GSE39828	1	Embelina	HEK 293T	GPL6480
GSE43166	4	Folato y vitamina B12	CHUB-S7	GPL14550
GSE44290	3	Extracto de hojas de Achyranthes aspera	PaCa-2	GPL4133
GSE45357	1	Gliadina	Caco2	GPL4133
GSE45804	1	Ácido lacciaco A	MCF7	GPL6244
GSE48668	3	Tocotrienoles	HeLa	GPL7020
GSE50945	2	Aceites de pescado y soja	Células de placenta	GPL10558
GSE50994	3	Silimarina	Huh7.5.1	GPL6244
GSE53049	2	Witaferina A	MCF7, MDA	GPL10558
GSE5556	3	Extractos de arándano, uva y vino rojo	Células de endotelio aórtico	GPL570
GSE55897	6	Indol-3-carbinol	MCF7, MDA, T47D, ZR75	GPL10904
GSE56245	1	Galato de epigalocatequina	MCF7	GPL13252
GSE56265	3	Ácido lisofosfatídico	MCF7. MDA, PC3	GPL570
GSE56496	3	Extracto de romero	SW620	GPL6480
GSE58749	8	Ácido retinoico y ácido 3,4-didehidroretinoico	Queratinocitos	GPL17692
GSE65012	2	Cortistatina A	K562, MOLM-14	GPL570
GSE65014	1	Cortistatina A	MOLM-14	GPL570
GSE65019	1	Cortistatina A	MV4-11	GPL570
GSE65030	1	Cortistatina A	HCT116	GPL11154
GSE65397	6	Ácidos grados de yema de huevo	MCF7	GPL17077
GSE65527	1	Urolitina A	LNCaP	GPL13667
GSE65722	2	Extractos de romero y ácido carnósico	HT29	GPL16686
GSE71606	1	Extracto de hoja de Tamarindus indica	HepG2	GPL6244
GSE71717	12	Genisteina	Células Ishikawa	GPL570
GSE7259	2	Quercetina	Caco2	GPL571
GSE73650	1	Transresveratrol	Caco2	GPL16686

ID GEO	Experimentos	Alimentos y compuestos bioactivos	Líneas celulares	Plataforma
GSE74212	1	Eusynstyelamida B	LNCaP	GPL16604
GSE75181	1	Extracto de té verde	Condrocitos	GPL10558
GSE79473	1	Extracto de hoja de Barringtonia racemosa	HepG2	GPL6244
GSE81277	2	6-acetoxi-anopterina, vinblastina	LNCaP	GPL16604
GSE83893	2	Cebolla blanca y amarilla	Caco2	GPL11532
GSE85871	77	Ingredientes de medicina tradicional china	MCF7	GPL571
GSE86044	1	Englerina A	A498	GPL17077
GSE86045	1	Englerina A	A498	GPL17077
GSE86046	1	Englerina A	A498	GPL17077
GSE94548	3	Extracto de tomate	MCF7	GPL15207
GSE9647	2	Extracto de manzana	HUVEC	GPL570

2. INTERFAZ GRÁFICA

La aplicación web de la plataforma de minería de datos ha sido desarrollada utilizando el marco de trabajo “Ruby on Rails”, basado en el paradigma del patrón Modelo Vista Controlador (MVC). Para el diseño de la interfaz gráfica del usuario, se ha utilizado la librería de código abierto “Bootstrap”, la cual proporciona plantillas, formularios, botones, cuadros, menús de navegación y otros elementos prediseñados, basados en lenguajes HTML y CSS, así como extensiones de JavaScript adicionales. Es compatible con la mayoría de navegadores web, y responde a las visualizaciones realizadas desde un terminal móvil.

3. MÓDULO EXPLORATORIO

Para el módulo exploratorio de la plataforma de minería de datos en nutrigenómica, se ha creado una nueva base de datos MySQL en el servidor web, a partir de la base de datos actualizada. Esta base de datos puede ser consultada utilizando símbolos de genes humanos. Las consultas devuelven los correspondientes niveles de expresión en los experimentos de nutrigenómica, además de información estadística obtenida tras el análisis de expresión diferencial. Los resultados de las consultas se presentan en una tabla interactiva, implementada gracias al complemento “DataTables” del lenguaje Javascript, y puede ser descargada en formatos Excel o PDF. El módulo exploratorio de la aplicación también permite generar dos tipos de visualizaciones interactivas: a) un gráfico de líneas, que permite explorar el nivel de expresión de los genes interrogados de manera simultánea en los experimentos de nutrigenómica incluidos, implementada gracias a la librería de gráficos “Plotly’s” (en código Python), b) un mapa de calor que permite agrupar los experimentos incluidos en la base de datos, y buscar patrones a partir de los valores de expresión diferencial de un conjunto de genes, implementado gracias a la librería de Javascript “Clustergrammer” (76).

4. MÓDULO ANALÍTICO

Para el módulo analítico de la plataforma de minería de datos en nutrigenómica, se ha utilizado la base de datos actualizada que contiene las firmas moleculares que definen los experimentos de nutrigenómica. Esta base de datos ha sido convertida al formato GMT, requerido por el algoritmo GSEA, para permitir la comparación de las firmas moleculares. El módulo analítico permite comparar un perfil de expresión diferencial externo, con las firmas moleculares incluidas en la base de datos, gracias a la implementación del algoritmo GSEA versión 2.2.2. El algoritmo está configurado para usar un sistema de puntuación ponderado (parámetro "scoring scheme" igual a "weighted"), y utiliza la modalidad "pre-ranked" para su ejecución, ya que los datos introducidos para el análisis son el resultado de un análisis de expresión diferencial. Las consultas enviadas se guardan en el servidor con un identificador único, y tras finalizar el análisis, se mantienen accesibles durante al menos un mes. Los resultados se presentan en una tabla interactiva implementada gracias al complemento "DataTables" del lenguaje Javascript, y que puede ser descargada en formatos Excel o PDF.

5. ENRIQUECIMIENTO DE FUNCIONES MOLECULARES

Desde el módulo analítico se permite realizar un enriquecimiento de las funciones moleculares sobre representadas a partir de los genes coincidentes entre un perfil de expresión diferencial externo y las firmas moleculares de la base de datos. Para esta tarea se ha implementado un servicio web ("webservice") que permite enviar los datos al servidor de Panther (74) para su análisis funcional. El servicio web devuelve un archivo de texto con los resultados del enriquecimiento funcional, obtenidos tras realizar el test exacto de Fisher. Estos resultados incluyen una línea por cada gen analizado, junto con la correspondiente función molecular asignada, los valores estadísticos de valor *p* y su corrección por el método de Benjamini-Hochberg (FDR) (52).

RESULTADOS

CAPÍTULO 1: ANÁLISIS EXPERIMENTAL DEL EFECTO DEL HIDROXITIROSOL EN LA EXPRESIÓN DE MICRO ARN'S EN EL HÍGADO DE RATÓN

1. EL HIDROXITIROSOL REGULA LA EXPRESIÓN DE MIARN'S EN EL HÍGADO DE RATÓN

Las muestras de hígado de ratón obtenidas tras el experimento han sido secuenciadas y se ha analizado el perfil de expresión de los miARN's. En la muestra se ha detectado la presencia de un total de 247 miARN's en su forma madura. Tras el análisis de expresión diferencial de los datos de secuenciación, se han obtenido un total de 4 miARN's diferencialmente expresados en el grupo de ratones que consumió la dieta suplementada en HT, con una significancia estadística por debajo del umbral de 0.05 tras el ajuste por testeo múltiple (**Tabla 5**). De estos, los miARN's mmu-miR-802-5p, mmu-miR-30a-5p y mmu-miR-146b han sido sobreexpresados, mientras que únicamente el miARN mmu-miR-423-3p ha sido reprimido. El nivel de expresión de estos miARN's es relativamente alto, por encima de 6 cuentas por millón en escala logarítmica en base 2 (columna logCPM), lo que corresponde a una media de más de 64 moléculas de miARN por cada millón de lecturas de secuenciación obtenidas en todas las muestras (**Figura 17**).

Tabla 5: Resultados del análisis de expresión diferencial de miARN's en hígado de ratón tras el consumo de una dieta suplementada con HT

miARN	logFC	logCPM	PValue	FDR
mmu-miR-802-5p	0.81959113	6.5680416	0.00017826	0.02628041
mmu-miR-423-3p	-0.50835122	7.05571795	0.0002128	0.02628041
mmu-miR-30a-5p	0.52381304	15.6957709	0.00045553	0.03001062
mmu-miR-146b-5p	0.49501708	6.46556765	0.000486	0.03001062
mmu-miR-128-3p	-0.4939714	5.77610583	0.00259205	0.12804708
mmu-miR-30e-5p	0.36526651	11.7488764	0.00347326	0.14298262
mmu-miR-542-3p	0.40382169	7.8373455	0.0050216	0.17719073
mmu-miR-30d-5p	0.33633391	13.1587041	0.00749659	0.2314573
mmu-miR-101a-3p	0.28501555	14.0099992	0.0089362	0.24524911
mmu-miR-802-3p	0.61369055	3.71328813	0.01380621	0.34101347
mmu-miR-34a-5p	0.51623395	4.76664695	0.01631801	0.36641345
mmu-miR-98-5p	-0.31390096	8.17780846	0.02146216	0.44176271
mmu-miR-1249-3p	-1.14959595	1.57882736	0.02457152	0.46685886
mmu-miR-22-5p	-0.30284476	7.39071858	0.02674397	0.47184001
mmu-miR-744-5p	-0.31353127	6.28587342	0.03294848	0.48204424
mmu-miR-322-5p	-0.47529028	4.1210241	0.03550189	0.48204424
mmu-miR-148a-3p	0.37813958	18.1854355	0.03656669	0.48204424
mmu-let-7e-5p	0.30422454	8.66691578	0.037033	0.48204424
mmu-miR-149-5p	0.47169712	5.36638713	0.03708033	0.48204424
mmu-miR-674-3p	0.32266195	5.80383911	0.0414814	0.50705706
mmu-miR-339-5p	-0.34654744	6.00138383	0.04311012	0.50705706

miARN	logFC	logCPM	PValue	FDR
mmu-miR-140-3p	0.23900843	7.88065486	0.04546117	0.51040499
mmu-miR-1198-5p	-0.32523616	5.26647239	0.05541572	0.58520177
mmu-miR-30e-3p	-0.2144278	8.97618072	0.05686171	0.58520177
mmu-miR-31-5p	0.27785308	6.5091211	0.06140851	0.60671603
mmu-miR-32-3p	-0.66611454	2.55215857	0.0815272	0.72561989
mmu-miR-30d-3p	-0.28589748	5.02636157	0.09051825	0.72561989
mmu-miR-122-3p	-0.26026678	8.35151931	0.09161218	0.72561989
mmu-miR-532-3p	-0.55552718	2.72662593	0.09514211	0.72561989
mmu-miR-1968-5p	-0.60836513	2.61425817	0.0959571	0.72561989
mmu-miR-22-3p	-0.20109296	12.1652151	0.09602438	0.72561989
mmu-miR-139-5p	-0.23073954	6.61283012	0.0964566	0.72561989
mmu-let-7e-3p	0.52875689	3.36839047	0.09694517	0.72561989
mmu-miR-106b-3p	-0.28152474	5.40804422	0.10295053	0.74563417
mmu-miR-381-3p	0.40444072	3.69771159	0.11490244	0.74563417
mmu-miR-421-3p	-0.50954185	2.4994067	0.12170753	0.74563417
mmu-miR-30c-5p	-0.17706385	12.4175185	0.12387659	0.74563417
mmu-let-7d-5p	-0.17014012	10.4077067	0.12416254	0.74563417
mmu-miR-146a-5p	0.20432583	11.6016255	0.12545622	0.74563417
mmu-miR-1843b-5p	-0.23179006	6.54578692	0.12798589	0.74563417
mmu-miR-1839-3p	0.42459541	3.30800153	0.13029771	0.74563417
mmu-miR-872-3p	-0.45139987	2.77827094	0.13190213	0.74563417
mmu-miR-320-3p	0.24724428	5.30978612	0.13263013	0.74563417
mmu-let-7d-3p	-0.2219155	6.85330447	0.1351025	0.74563417
mmu-miR-152-5p	-0.25593104	5.34070721	0.13603791	0.74563417
mmu-miR-335-5p	0.23161029	7.20139614	0.13916373	0.74563417
mmu-miR-300-3p	0.53536373	2.45865421	0.14274486	0.74563417
mmu-miR-143-3p	0.19322232	13.7353169	0.14553331	0.74563417
mmu-miR-465c-5p	-0.61163109	2.19968548	0.14791933	0.74563417
mmu-miR-107-3p	-0.17841736	8.78305938	0.15864611	0.78371177
mmu-miR-200b-3p	0.20581966	8.78508706	0.17407024	0.82570124
mmu-miR-871-3p	-0.58643761	2.48356346	0.17611199	0.82570124
mmu-miR-378a-5p	-0.22387483	5.43166436	0.17717476	0.82570124
mmu-miR-16-1-3p	-0.2522856	4.56105037	0.18690411	0.84083751
mmu-miR-881-3p	0.2940737	5.39073804	0.18723103	0.84083751
mmu-miR-16-5p	0.14850009	10.024345	0.19770061	0.86560276
mmu-miR-676-5p	-0.32983214	3.99713771	0.19975448	0.86560276
mmu-miR-425-5p	0.19012962	5.31395969	0.2123165	0.89283577
mmu-let-7f-5p	-0.15983813	15.4765342	0.21326846	0.89283577
mmu-miR-339-3p	-0.38462355	3.42312512	0.22591001	0.92497891
mmu-let-7a-5p	-0.12172809	13.5250702	0.23027296	0.92497891
mmu-miR-200a-5p	0.28421492	3.71233372	0.23291673	0.92497891
mmu-miR-10a-5p	0.12030849	13.2114202	0.23897242	0.92497891
mmu-miR-145a-3p	-0.15543531	7.93603073	0.23967065	0.92497891

miARN	logFC	logCPM	PValue	FDR
mmu-miR-24-2-5p	0.16650682	7.13183817	0.25158056	0.93341745
mmu-miR-338-3p	0.47332424	2.44537962	0.26552892	0.93341745
mmu-miR-127-3p	0.19244216	5.03180814	0.26639271	0.93341745
mmu-miR-379-5p	0.2253973	4.55276835	0.26805377	0.93341745
mmu-miR-126a-3p	0.1102169	14.4963669	0.26840738	0.93341745
mmu-miR-744-3p	-0.61075776	1.88604744	0.26857492	0.93341745
mmu-miR-183-5p	-0.23490411	6.89433518	0.26983679	0.93341745
mmu-miR-1964-3p	-0.42007134	3.03482604	0.27209008	0.93341745
mmu-miR-221-5p	0.18446799	6.40553724	0.278737	0.93341745
mmu-miR-212-5p	0.38875444	2.25742832	0.28465943	0.93341745
mmu-miR-322-3p	0.13722536	6.81811796	0.28812093	0.93341745
mmu-miR-215-5p	-0.41037851	4.96616679	0.29186557	0.93341745
mmu-miR-1981-5p	0.24558004	4.43810533	0.29633505	0.93341745
mmu-miR-27a-5p	0.2503361	4.1936352	0.30189027	0.93341745
mmu-miR-532-5p	0.12352863	8.57394819	0.30311042	0.93341745
mmu-miR-34c-5p	-0.3190183	3.84614784	0.30957287	0.93341745
mmu-miR-342-3p	0.22117823	4.71412788	0.30975782	0.93341745
mmu-miR-199b-5p	-0.24200891	4.77151244	0.31577288	0.93341745
mmu-miR-30a-3p	-0.10839034	10.5408027	0.3166803	0.93341745
mmu-miR-186-5p	-0.13433435	8.18789127	0.31896609	0.93341745
mmu-miR-26a-2-3p	-0.31205268	3.74448825	0.32121653	0.93341745
mmu-miR-26a-5p	-0.10741447	14.7637575	0.32968739	0.93969936
mmu-miR-20a-5p	-0.11815077	9.16092756	0.33224713	0.93969936
mmu-miR-378a-3p	0.15662375	12.2907379	0.3379758	0.93969936
mmu-miR-99a-5p	-0.11621289	12.3082925	0.33907386	0.93969936
mmu-miR-340-5p	0.12533335	12.3250649	0.34524299	0.93969936
mmu-miR-192-3p	-0.23372235	3.3095141	0.34620503	0.93969936
mmu-miR-1948-3p	0.41198714	2.18485815	0.35064373	0.94140219
mmu-miR-223-3p	0.2922462	3.6549346	0.3612596	0.94417387
mmu-miR-192-5p	0.11576198	16.3267728	0.36150824	0.94417387
mmu-miR-221-3p	-0.10030702	9.2257185	0.38014771	0.94417387
mmu-miR-182-5p	-0.22665747	8.6983768	0.38858636	0.94417387
mmu-miR-19b-3p	0.10802811	9.41272494	0.39359923	0.94417387
mmu-miR-200c-3p	-0.15794666	4.94449978	0.39419682	0.94417387
mmu-miR-338-5p	0.25123056	3.36909278	0.39893528	0.94417387
mmu-miR-199a-5p	-0.11239814	9.32947003	0.40661542	0.94417387
mmu-let-7g-5p	-0.08381189	15.3182675	0.41008069	0.94417387
mmu-miR-511-3p	0.22709106	4.50133579	0.41332836	0.94417387
mmu-miR-23a-3p	0.14204034	8.38954869	0.41643566	0.94417387
mmu-miR-30c-2-3p	0.1204631	8.46940814	0.41792374	0.94417387
mmu-miR-125b-2-3p	0.13659764	5.89696433	0.42182648	0.94417387
mmu-miR-7a-5p	0.13839367	11.2987783	0.42292629	0.94417387
mmu-miR-148b-5p	-0.18260882	4.68763055	0.42362984	0.94417387

miARN	logFC	logCPM	PValue	FDR
mmu-miR-503-5p	-0.30629881	2.34597679	0.43084628	0.94417387
mmu-miR-374b-5p	-0.11180139	8.24075018	0.43105203	0.94417387
mmu-miR-351-5p	-0.12155001	5.91650288	0.43469662	0.94417387
mmu-miR-152-3p	0.09342393	11.0088812	0.44216678	0.94417387
mmu-miR-92a-3p	-0.1594087	4.27797481	0.44303619	0.94417387
mmu-miR-17-5p	-0.10644751	7.14477387	0.44469956	0.94417387
mmu-miR-25-3p	0.0970982	9.97775347	0.44527073	0.94417387
mmu-miR-148a-5p	0.10902271	8.4423712	0.4554257	0.94417387
mmu-miR-210-3p	0.25388277	3.12552133	0.45944943	0.94417387
mmu-miR-195a-3p	0.16878471	4.04404597	0.46050289	0.94417387
mmu-miR-148b-3p	-0.08677072	10.1816257	0.46209639	0.94417387
mmu-miR-365-3p	-0.11660997	6.29993864	0.46244681	0.94417387
mmu-miR-193a-3p	0.28679907	3.13922471	0.47042627	0.94417387
mmu-miR-30c-1-3p	-0.20482278	4.35880962	0.4721362	0.94417387
mmu-miR-1839-5p	0.10343641	8.8963961	0.47774706	0.94417387
mmu-miR-99b-3p	0.18354227	3.73307079	0.48577208	0.94417387
mmu-miR-7a-1-3p	0.15625356	3.71447987	0.49050188	0.94417387
mmu-miR-33-5p	-0.28850925	2.32895184	0.49071354	0.94417387
mmu-miR-151-5p	0.1204758	5.69604233	0.49199565	0.94417387
mmu-miR-27b-5p	0.35017272	2.11540692	0.49567105	0.94417387
mmu-miR-132-3p	0.24032187	2.97870028	0.49598401	0.94417387
mmu-miR-125b-1-3p	0.28614039	2.58171311	0.49705392	0.94417387
mmu-miR-150-5p	-0.21183891	4.15623624	0.5015302	0.94417387
mmu-miR-181d-5p	0.12894919	5.163798	0.50531191	0.94417387
mmu-miR-129-5p	-0.21070516	2.6480323	0.50754706	0.94417387
mmu-miR-125a-3p	0.27993814	2.60854507	0.50840132	0.94417387
mmu-miR-144-5p	-0.4368909	3.82316543	0.5161175	0.94897358
mmu-miR-140-5p	-0.10554326	6.89572767	0.51866977	0.94897358
mmu-miR-30b-3p	0.11001317	5.46997664	0.52393218	0.95155329
mmu-miR-194-5p	0.06144612	13.9413077	0.53374334	0.95854209
mmu-let-7b-5p	-0.06607083	11.3703808	0.55213772	0.95854209
mmu-miR-191-3p	0.20157471	2.39506499	0.55448597	0.95854209
mmu-miR-130a-3p	0.12279878	5.6249071	0.55498828	0.95854209
mmu-miR-185-5p	0.07407506	9.56114978	0.55656237	0.95854209
mmu-miR-1843a-3p	-0.148762	4.63938478	0.55786085	0.95854209
mmu-miR-18a-5p	-0.24208666	1.98753162	0.55862609	0.95854209
mmu-miR-1a-3p	-0.12683065	6.8619257	0.55882616	0.95854209
mmu-miR-181b-5p	-0.13827441	5.50608122	0.56455498	0.9616902
mmu-miR-218-5p	0.17646228	2.97043228	0.57108286	0.96614703
mmu-miR-19a-3p	0.08976646	7.11118318	0.58334304	0.96865412
mmu-miR-21a-3p	-0.13023855	3.7919618	0.58742282	0.96865412
mmu-miR-9-5p	0.3247028	2.42175456	0.58822011	0.96865412
mmu-miR-361-5p	-0.10491028	5.0771021	0.58987868	0.96865412

miARN	logFC	logCPM	PValue	FDR
mmu-miR-676-3p	-0.18732794	4.00831825	0.59679699	0.96865412
mmu-miR-700-3p	0.29141844	2.17053712	0.60061372	0.96865412
mmu-miR-23b-3p	0.06518607	8.45944397	0.60788521	0.96865412
mmu-miR-450b-5p	0.11521185	2.81881781	0.61053159	0.96865412
mmu-let-7f-2-3p	0.27882908	2.31355589	0.61095519	0.96865412
mmu-miR-181a-2-3p	-0.16041799	2.54538167	0.61189066	0.96865412
mmu-miR-434-5p	0.23211529	2.9931904	0.61688171	0.96865412
mmu-miR-10a-3p	0.10044994	6.14417169	0.61962491	0.96865412
mmu-miR-574-5p	0.2258335	3.39644921	0.62691371	0.96928166
mmu-miR-361-3p	-0.0594319	7.72056266	0.62787476	0.96928166
mmu-miR-125b-5p	-0.05421286	9.6446944	0.63679447	0.97477837
mmu-miR-26b-5p	-0.05051586	10.8158753	0.65089146	0.97477837
mmu-miR-28a-5p	-0.07718497	6.06342608	0.65167002	0.97477837
mmu-miR-143-5p	-0.13309885	4.07602332	0.65353883	0.97477837
mmu-miR-29b-2-5p	-0.21272547	2.09670493	0.65369817	0.97477837
mmu-miR-222-3p	-0.10184726	4.58951979	0.65675628	0.97477837
mmu-miR-350-3p	0.17842313	3.47175955	0.65906068	0.97477837
mmu-miR-96-5p	-0.12997581	4.93325867	0.6779357	0.98024017
mmu-miR-467a-5p	0.08522918	3.1494788	0.68546065	0.98024017
mmu-miR-365-2-5p	0.21901274	2.32244135	0.6858488	0.98024017
mmu-miR-423-5p	-0.06038923	7.33441239	0.6874697	0.98024017
mmu-miR-674-5p	0.14981554	2.34373507	0.68999066	0.98024017
mmu-miR-99b-5p	0.0556976	9.40797112	0.69197501	0.98024017
mmu-miR-29a-3p	-0.04385248	11.6309869	0.69688089	0.98024017
mmu-miR-741-3p	0.14964032	3.2592566	0.69722834	0.98024017
mmu-miR-652-3p	-0.13656354	3.54949097	0.70223444	0.98024017
mmu-miR-106b-5p	0.09620539	5.36800147	0.70243931	0.98024017
mmu-let-7b-3p	-0.06057318	4.07751381	0.70755627	0.98183369
mmu-miR-450a-5p	-0.05703414	6.3970552	0.72173954	0.98889851
mmu-miR-194-2-3p	-0.06383028	6.64648266	0.72411743	0.98889851
mmu-miR-362-3p	0.08750242	5.32660849	0.72465842	0.98889851
mmu-miR-145a-5p	0.04657966	9.49803327	0.74042054	0.990371
mmu-miR-24-1-5p	0.1143087	3.29665728	0.74900945	0.990371
mmu-miR-501-3p	-0.07203318	4.19706351	0.75195052	0.990371
mmu-miR-29b-3p	0.05981473	6.83838055	0.75532722	0.990371
mmu-miR-465b-5p	-0.20647282	2.27668173	0.75960893	0.990371
mmu-miR-671-3p	-0.05716552	4.05543167	0.76627824	0.990371
mmu-miR-27a-3p	0.0438367	10.7629906	0.76728724	0.990371
mmu-miR-330-5p	-0.13477401	2.72335224	0.77009651	0.990371
mmu-miR-503-3p	0.04362625	2.53395758	0.77434897	0.990371
mmu-miR-155-5p	0.06270162	4.35867121	0.77914054	0.990371
mmu-miR-28a-3p	-0.03662573	6.46511086	0.78412666	0.990371
mmu-miR-592-5p	-0.07632633	4.62706301	0.78669818	0.990371

miARN	logFC	logCPM	PValue	FDR
mmu-miR-32-5p	-0.03113582	6.70351801	0.79055594	0.990371
mmu-miR-24-3p	0.03087316	9.95982945	0.79995049	0.990371
mmu-let-7c-5p	0.02690637	13.762667	0.80051297	0.990371
mmu-miR-195a-5p	0.04500837	5.89478813	0.80653484	0.990371
mmu-miR-214-3p	-0.06027546	3.89612993	0.81173819	0.990371
mmu-miR-21a-5p	0.026773	16.1406	0.81251677	0.990371
mmu-miR-122-5p	-0.02247956	17.3660803	0.81518605	0.990371
mmu-miR-598-3p	0.05667156	4.15675259	0.81532786	0.990371
mmu-miR-31-3p	0.1097595	2.10994275	0.8178322	0.990371
mmu-miR-7b-5p	0.02734372	11.1889161	0.82109484	0.990371
mmu-miR-328-3p	-0.04343861	6.81893948	0.82282966	0.990371
mmu-miR-200a-3p	0.03645291	8.61708402	0.8307562	0.990371
mmu-miR-3105-3p	0.11995086	2.85109424	0.83259939	0.990371
mmu-miR-191-5p	-0.02521558	10.9321779	0.83775243	0.990371
mmu-miR-137-3p	-0.08091214	2.92009288	0.84376461	0.990371
mmu-miR-15a-5p	0.06772862	4.50451075	0.85355313	0.990371
mmu-miR-214-5p	-0.00855333	3.31997765	0.85491501	0.990371
mmu-miR-15b-5p	0.02606256	5.35042447	0.85544731	0.990371
mmu-miR-101b-3p	0.0267991	9.67368851	0.86050956	0.990371
mmu-miR-193a-5p	0.05436714	3.59706832	0.863442	0.990371
mmu-miR-181c-5p	0.06388414	2.52072902	0.86435419	0.990371
mmu-miR-93-5p	-0.02261993	8.1769257	0.8734065	0.990371
mmu-miR-144-3p	-0.18957302	2.84944596	0.87346519	0.990371
mmu-miR-541-5p	0.05313828	3.60288451	0.87689953	0.990371
mmu-miR-335-3p	0.03355331	6.08120615	0.87812311	0.990371
mmu-miR-26b-3p	-0.0802028	3.07153063	0.88195801	0.990371
mmu-miR-126a-5p	0.01765726	10.4457022	0.89198295	0.990371
mmu-miR-411-5p	0.03766637	3.44507186	0.90054354	0.990371
mmu-miR-326-3p	-0.15619317	1.95821521	0.90540806	0.990371
mmu-miR-151-3p	-0.0234361	7.60910462	0.90578086	0.990371
mmu-miR-362-5p	0.06613426	4.12180118	0.91482198	0.990371
mmu-miR-30b-5p	-0.00966752	9.78767588	0.91567993	0.990371
mmu-miR-27b-3p	-0.01118715	13.7254075	0.91675913	0.990371
mmu-miR-345-3p	0.03646796	3.80465909	0.91873853	0.990371
mmu-miR-132-5p	0.11944281	2.28558486	0.92102248	0.990371
mmu-miR-434-3p	-0.08627846	2.24978586	0.92192919	0.990371
mmu-miR-103-3p	-0.01299311	10.8404121	0.92910984	0.990371
mmu-miR-1843a-5p	-0.01204075	8.4371152	0.93119567	0.990371
mmu-miR-872-5p	-0.02107919	6.98228801	0.93183116	0.990371
mmu-miR-181a-1-3p	0.03403461	2.93059663	0.93742962	0.990371
mmu-let-7c-1-3p	-0.0057653	2.71822897	0.93824621	0.990371
mmu-miR-429-3p	-0.02943094	5.18351672	0.94803791	0.99313849
mmu-miR-340-3p	-0.01319975	5.53508787	0.94890965	0.99313849

miARN	logFC	logCPM	PValue	FDR
mmu-miR-100-5p	0.00560535	10.6113969	0.96324543	0.99627033
mmu-let-7f-1-3p	-0.00173731	4.23796149	0.96780372	0.99627033
mmu-miR-1948-5p	-0.01132494	4.38059608	0.96855111	0.99627033
mmu-miR-574-3p	0.00936574	4.46989171	0.96942239	0.99627033
mmu-let-7i-5p	-0.00321063	12.5778685	0.9743263	0.99627033
mmu-miR-181a-5p	-0.00345197	7.84653946	0.97610291	0.99627033
mmu-miR-125a-5p	-0.00271315	8.96098871	0.99285935	1
mmu-miR-187-3p	0.0067814	4.86971879	1	1
mmu-miR-223-5p	-0.07505897	2.82218553	1	1
mmu-miR-181c-3p	-0.03693757	2.85155237	1	1
mmu-miR-15b-3p	-0.01736502	3.21432569	1	1

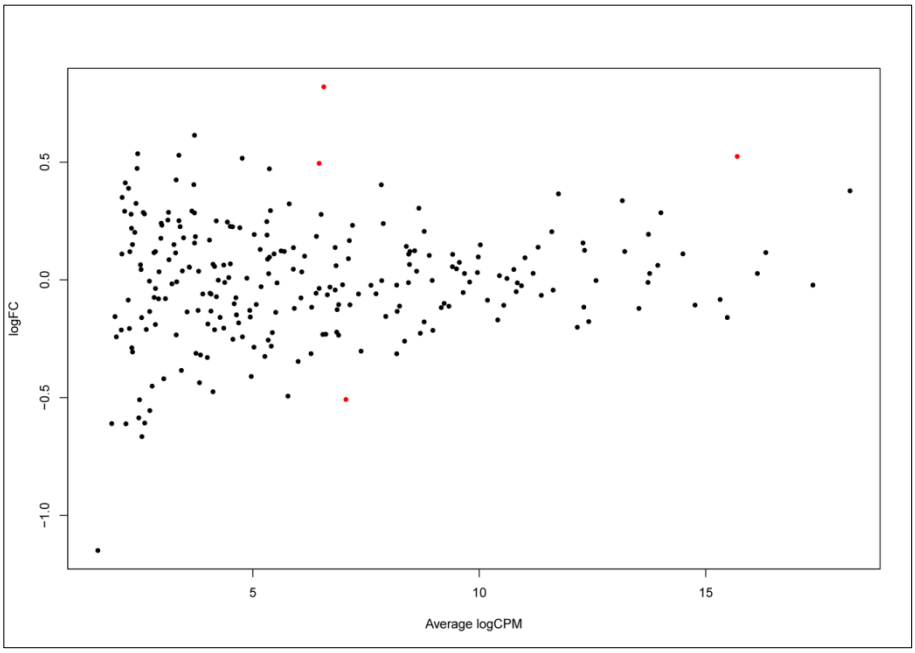


Figura 17: Gráfica de expresión relativa y media (MA plot) que muestra el nivel de expresión diferencial entre ambos grupos junto a la media del nivel de expresión de los miARN's analizados. En rojo aparecen los miARN's diferencialmente expresados.

2. Red de interacción miARN–ARNm

Debido a su actividad reguladora post-transcripcional, un pequeño cambio en los niveles de expresión de un miARN puede implicar un importante efecto biológico. Con el fin de identificar las implicaciones funcionales de los resultados obtenidos, se ha generado una red de interacción miARN-ARNm, integrando los potenciales genes diana de los miARN's diferencialmente expresados entre ambas condiciones. La red de interacción contiene un total de 279 genes, potencialmente

regulados por al menos 2 de los 4 miARN's diferencialmente expresados (**Figura 18**). De entre estos, se han identificado 12 genes como potencial diana común de los 4 miARN's diferencialmente expresados (**Tabla 6**).

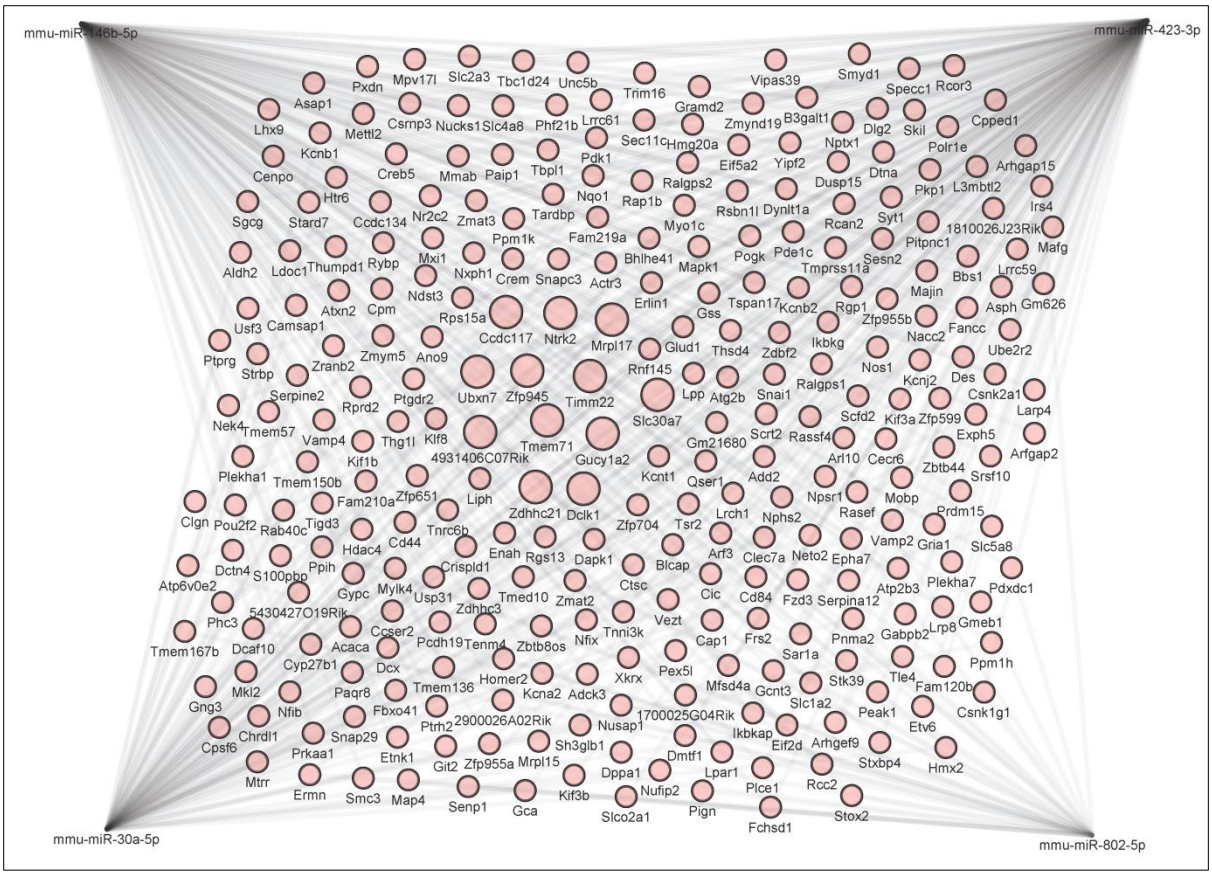


Figura 18: Red de interacción entre los miARN's diferencialmente expresados tras el experimento junto con sus potenciales genes dianas. El tamaño de los puntos de cada gen correlaciona directamente con el número de interacciones con los miARN's.

Tabla 6: Lista y descripción de los 12 genes diana de los 4 miARN's diferencialmente expresados.

Simbolo de gen	Descripción
Ccdc117	Proteína 117 con dominio en espiral
Ntrk2	Receptor neurotrófico tirosina quinasa tipo 2
Mrpl17	Proteína ribosomal de la mitocondria L17
Timm22	Translocasa 22 de la membrana interna de la mitocondria
Zfp945	Proteína con dedo de zinc 945
Ubxn7	Proteína 7 con dominio UBX
Tmem71	Proteína transmembrana 71
Slc30a7	Miembro 7 de la familia de transportadores de solutos (transportador de zinc)
Gucy1a2	Ciclasa guanilato 1 soluble alfa 2
4931406C07Rik	ADNc RIKEN 4931406C07
Zdhhc21	Dedo de zinc 21 con dominio DHHC
Dclk1	Quinasa 1 similar a doblecortina

3. ANÁLISIS FUNCIONAL DE LOS MICRO ARN'S IDENTIFICADOS

Finalmente, se ha realizado un análisis de enriquecimiento funcional a partir de los 279 genes incluidos en la red de interacción, utilizando la base de datos Panther, para identificar vías metabólicas potencialmente sobre-representadas en las cuales participan los genes diana identificados. El resultado muestra que los miARN's regulados por el consumo de una dieta suplementada en HT en el hígado de ratón están implicados en importantes vías metabólicas tales como la vía de señalización Wnt (P00057), la vía de señalización por el receptor de colecistoquinina B (CCKR) (P06959) y la vía de señalización de inflamación a través de citoquinas y quimiocinas (P00031) (**Figura 19**).

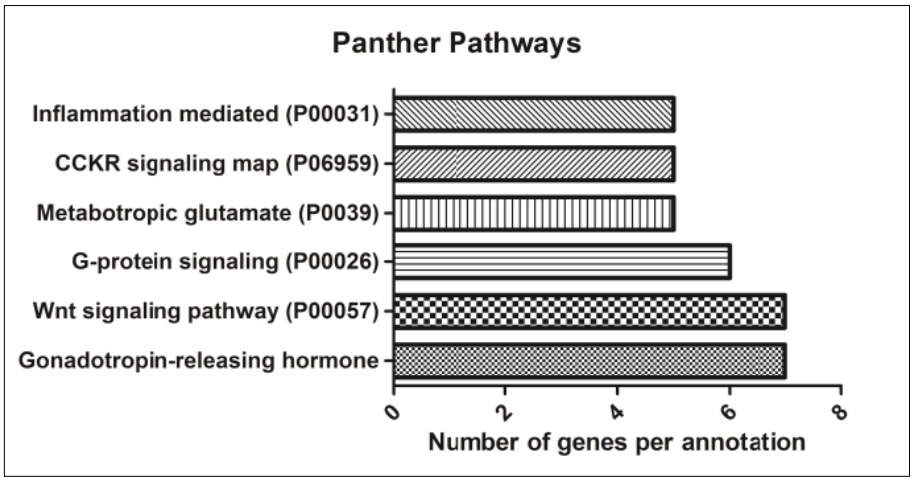


Figura 19: Resultados del enriquecimiento funcional de vías metabólicas entre los 279 genes incluidos en la red de interacción miARN- ARNm. Cada vía metabólica cuenta con al menos 5 genes anotados en ella.

CAPÍTULO 2: CREACIÓN Y ANÁLISIS INTEGRATIVO DE UNA BASE DE DATOS A PARTIR DE EXPERIMENTOS DE NUTRIGENÓMICA EN CÉLULAS HUMANAS.

1. AGRUPAMIENTO JERÁRQUICO DE EXPERIMENTOS COMPLETOS DE NUTRIGENÓMICA

El flujo de trabajo para el análisis integrativo se encuentra resumido en la **figura 20**. Utilizando la primera base de datos, obtenida a partir de las búsquedas realizadas hasta Enero de 2017, se ha realizado un análisis de agrupamiento jerárquico de los experimentos utilizando los valores de expresión diferencial log2 FC. La matriz de datos obtenida integra los resultados de hasta 18 arrays de expresión distintos, y por ello presenta la limitación de que sólo los genes comunes a todos los arrays de expresión han podido ser utilizados para el análisis integrativo. Para la integración de los datos de expresión de todos los arrays, se ha utilizado el identificador de gen correspondiente al identificador "Entrez". Efectivamente se da el caso de que dependiendo del array de expresión utilizado, el mismo gen puede estar identificado por símbolos distintos, debido a las actualizaciones de los símbolos de genes que ocurren constantemente en las bases de datos genómicas más utilizadas. El identificador "Entrez" se mantiene estable en la base de datos del "National Center for Biotechnology Information" (NCBI).

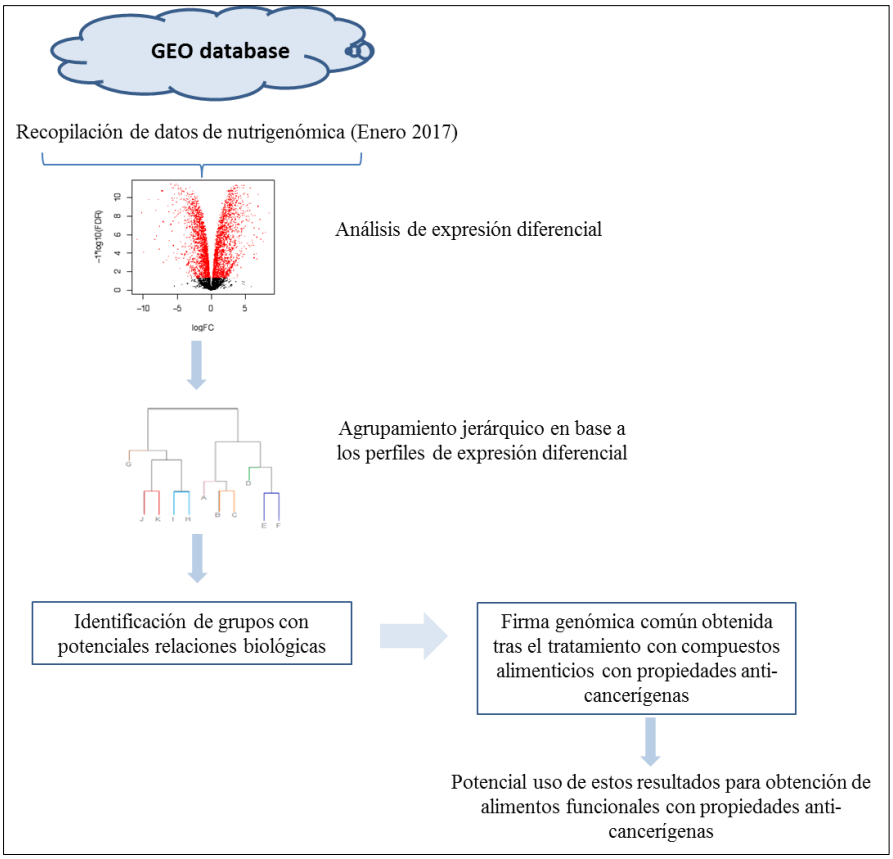


Figura 20: Esquema del flujo de trabajo para el análisis integrativo de la primera versión de la base de datos de experimentos de nutrigenómica.

Tras la integración de los datos de expresión de los genes comunes analizados en los distintos arrays de expresión, se ha obtenido una matriz con 4.659 genes presentes en 73 experimentos de nutrigenómica. El dendrograma resultante del análisis de agrupamiento jerárquico sobre estos datos corresponde a la **figura 21**. La inspección visual de este último revela que los experimentos tienden a agruparse de manera tejido-dependiente. Este resultado es coherente y se explica por la expresión de conjuntos de genes específicos a determinados tipos celulares. Por ejemplo, se han identificado:

- Agrupamientos de células HeLa tras su tratamiento con distintos Tocotrienoles (TCT) (alfa, beta y gama).

- Agrupamientos de células musculares (Miotubos) tras su tratamiento con distintos ácidos grasos (eicosapentaenoico, linoleico y oleico).

- Agrupamientos de adipocitos CHUB-S7 tras su tratamiento con distintas dosis de folato y vitamina D.

- Agrupamientos de células de placenta tras su tratamiento con aceite de pescado.

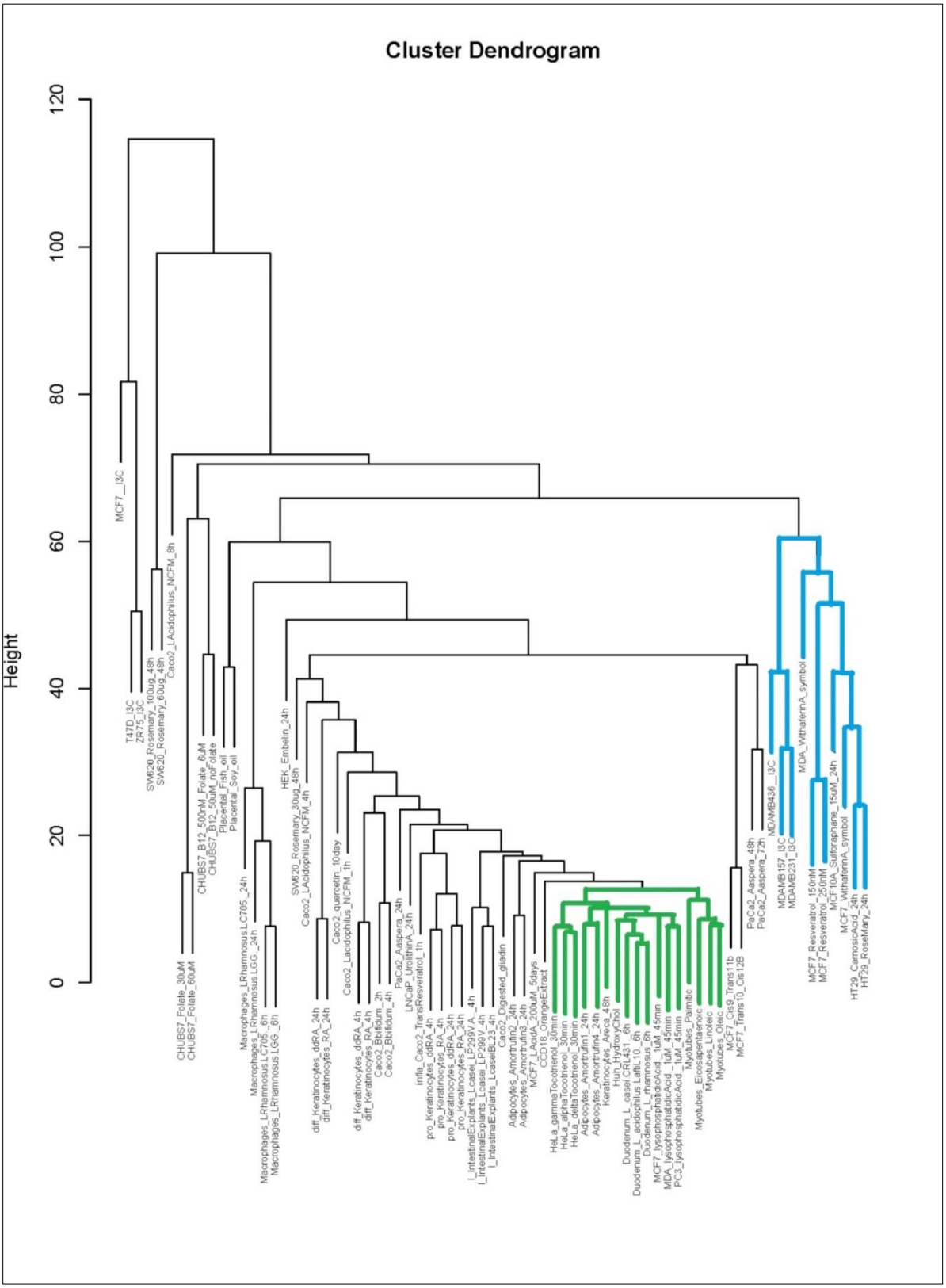


Figura 21: Dendrograma resultante del análisis de agrupamiento jerárquico a partir de los datos de expresión diferencial (log2 FC) obtenidos tras el análisis de cada experimento. Los dos grupos resaltados en colores azul y verde corresponden a grupos de experimentos tras el tratamiento con compuestos con potenciales propiedades anticancerígenas y antioxidantes/antiinflamatorias respectivamente.

Inspeccionando el dendrograma resultante del agrupamiento jerárquico, esta vez prestando atención al tipo de compuesto utilizado en el experimento, se han identificado relaciones potencialmente interesantes entre los distintos grupos. Existe un grupo relativamente grande (**Figura 21** en verde) que agrupa distintos tipos celulares tras su tratamiento con TCT (presente naturalmente en aceites vegetales), ácido lisofosfatídico (un lípido bioactivo con propiedades similares a las de los factores de crecimiento), hidroxicolesterol, dos distintos extractos de amorfrutinas (AMF) y tratamientos con diferentes cepas de Lactobacillus (LB). Paradójicamente, los 4 tratamientos con distintos extractos de AMF se agrupan en 2 grupos diferentes y relativamente distantes. Dado que los tratamientos realizados con estos extractos han utilizado la misma concentración de 30µM, este resultado evidencia la diferente actividad biológica que provocan respectivamente los tratamientos con los extractos de AMF 2-3 y 1-4.

La identificación del grupo coloreado en verde (**figura 21**), que agrupa distintos compuestos y tipos celulares, indica que estos compuestos podrían compartir propiedades biológicas, independientemente del tipo celular utilizado para el tratamiento. Efectivamente los compuestos incluidos en este grupo ya han demostrado previamente su capacidad antioxidante y antiinflamatoria (77-80). Con el objetivo de extraer información funcional sobre los efectos biológicos de estos compuestos, se han buscado coincidencias entre los 3.000 genes con mayor significancia estadística, obtenidos en cada uno de los resultados de expresión diferencial de los experimentos independientes incluidos en este grupo. Sin embargo las coincidencias son prácticamente inexistentes, lo cual impide profundizar en los mecanismos moleculares que pudieran compartir los tratamientos utilizando los compuestos con potencial actividad antioxidante y antiinflamatoria incluidos en este grupo.

2. AGRUPAMIENTO DE TRATAMIENTOS CON COMPUESTOS POTENCIALMENTE ANTICANCERÍGENOS

Continuando con la inspección visual del dendrograma (**Figura 21**), se ha identificado otro grupo de experimentos (coloreado en azul), el cual agrupa distintas células de origen cancerígeno (MCF7, HT29 y subtipos de MDA) expuestas a compuestos con potencial uso en alimentación, y que previamente han demostrado poseer propiedades anticancerígenas. Estos compuestos son el Indole-3-carbinol (I3C) (81), un extracto de romero (RSM), el ácido carnósico (CA), la witaferina A (WFNA) (82), el sulforafano (SFP) (83) y el resveratrol (RVT) (84). De este modo postulamos que estos compuestos podrían estar ejerciendo sus potenciales efectos anticancerígenos mediante un mecanismo o modo de acción en común. También es oportuno señalar que otros 2 experimentos que emplean el tratamiento con un extracto de RSM, a concentraciones de 60 y 100µg/mL

respectivamente, no se encuentran en el grupo identificado. Este hecho puede ser debido a las diferentes concentraciones de RSM utilizadas en los experimentos, las cuales son mayores que las del experimento con RSM incluido en el grupo coloreado en azul, realizado en células HT29 y utilizando una concentración de RSM de 30 µg/mL (los detalles del estudio son accesibles a través del identificador de GEO GSE65722). Las altas concentraciones de RSM utilizadas podrían ser demasiado elevadas e inducirían pronunciados niveles de muerte celular dirigidos por otros mecanismos moleculares.

3. DETERMINACIÓN DE LA INFLUENCIA DE LOS EFECTOS-LOTE ("BATCH- EFFECTS")

Los efectos-lote pueden influir en los patrones de agrupamiento observados en el dendrograma generado, el cual podría estar mostrando relaciones entre experimentos debidas a factores sin origen biológico, como por ejemplo experimentos generados en un mismo laboratorio, por un mismo técnico o incluso utilizando los mismos protocolos experimentales. Estos efectos aparecen cuando se analizan de manera integrativa distintos experimentos generados con arrays de expresión, con el objetivo de aumentar la robustez estadística de los resultados de expresión diferencial de genes en el estudio de enfermedades o tratamientos específicos, procedimiento conocido como meta-análisis. Generalmente en este tipo de análisis se utilizan los valores de expresión normalizados para cada muestra del experimento. Sin embargo, en el caso de este trabajo doctoral, cada experimento ha sido analizado de manera independiente, y se ha utilizado el valor de expresión diferencial para el agrupamiento jerárquico, con lo cual los resultados del agrupamiento jerárquico no deberían estar influenciados por los efectos-lote.

Para determinar la influencia de los efectos-lote en el dendrograma obtenido, se ha realizado el mismo proceso de integración de resultados, a partir de los mismos experimentos, utilizando esta vez el valor medio de expresión de cada gen en todos los arrays incluidos en el experimento (columna "AveExpr" del archivo de resultados, **tabla 2**), en lugar del valor log2 FC. Analizando la figura obtenida (**Figura 22**), en el nuevo dendrograma ya no es posible encontrar los grupos de experimentos identificados previamente en la **figura 21**. Además se aprecia claramente como todos los experimentos se encuentran agrupados estrictamente según el estudio de procedencia, ya que posiblemente fueron generados en el mismo laboratorio. Estas observaciones demuestran un pronunciado efecto-lote cuando se utilizan los valores medios de expresión génica para un análisis integrativo, en lugar de los valores de expresión diferencial log2 FC. De este modo, se puede concluir que el análisis de agrupamiento jerárquico realizado originalmente en este trabajo no sufre de tales efectos, y los grupos identificados pueden ser explicados por la presencia de modos de acción o mecanismos moleculares comunes a los compuestos incluidos.

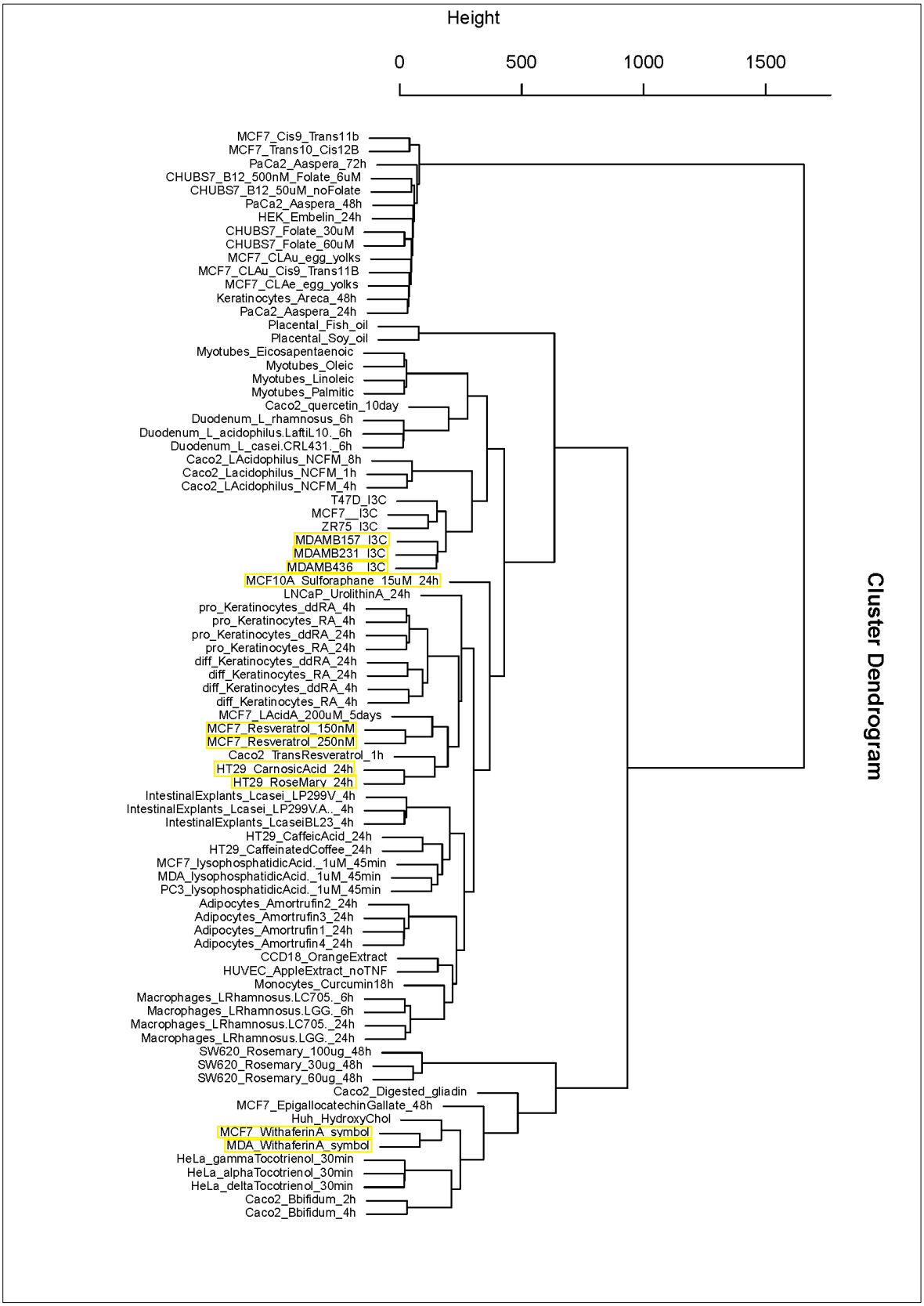


Figura 22: Dendrograma obtenido a partir del agrupamiento jerárquico de los experimentos de nutrigenómica utilizando el valor medio de expresión de cada gen en todos los arrays de cada experimento. El agrupamiento observado se da en función de cada estudio. En amarillo están señalados los experimentos que estaban incluidos en el grupo identificado (azul), que incluye compuestos con potenciales efectos anticancerígenos.

4. IDENTIFICACIÓN DE UNA FIRMA MOLECULAR DE 18 GENES CON PROPIEDADES ANTICANCERÍGENAS

Con el objetivo de identificar un mecanismo molecular común a los compuestos con potencial anticancerígeno identificados durante el agrupamiento jerárquico, se ha realizado una búsqueda de genes diferencialmente expresados en común en los experimentos identificados. Para ello se ha seleccionado un experimento de cada rama del grupo azul identificado en el agrupamiento jerárquico (**Figura 21**) como representante de cada compuesto con potenciales propiedades anticancerígenas. De esta manera se han seleccionado un total de 4 experimentos tras el tratamiento con 4 compuestos: CA, WFNA, SFP y RVT a 150nM. Con el fin de aumentar las posibilidades de encontrar coincidencias, y también debido a su alejamiento del grupo principal, los experimentos realizados con I3C no han sido incluidos en esta comparación. Tampoco se ha incluido el experimento realizado con RSM y células HT29, ya que naturalmente el CA se encuentra presente de manera abundante en el RSM. De hecho los resultados de ambos experimentos presentan una alta correlación al estar agrupados en una misma rama del dendrograma (**Figura 21**).

A partir de los resultados de expresión diferencial de estos 4 experimentos, analizados de manera independiente, se observa un importante solapamiento entre los 3.000 genes con mayor significancia estadística (**Figura 23**). De este modo, a partir de los 160 genes en común diferencialmente expresados entre los 4 experimentos, se ha realizado un análisis de enriquecimiento de los procesos biológicos en los cuales este conjunto de genes participa (**Tabla 7**). Los resultados de este análisis funcional muestran el término GO:0000278 (“mitotic cell cycle”) como el termino enriquecido más estadísticamente significativo, y engloba un total de 18 genes. Efectivamente este hallazgo tiene mucho sentido desde el punto de vista de una firma molecular con propiedades anticancerígenas, ya que afecta al ciclo de división celular. Además la inspección de los valores de expresión diferencial de los 18 genes seleccionados muestra un pronunciado patrón común de represión en todos los experimentos incluidos en el grupo de compuestos con potenciales propiedades anticancerígenas (**Tabla 8**). La lista de genes incluidos en la firma molecular identificada y su descripción se encuentra disponible en la **tabla 9**.

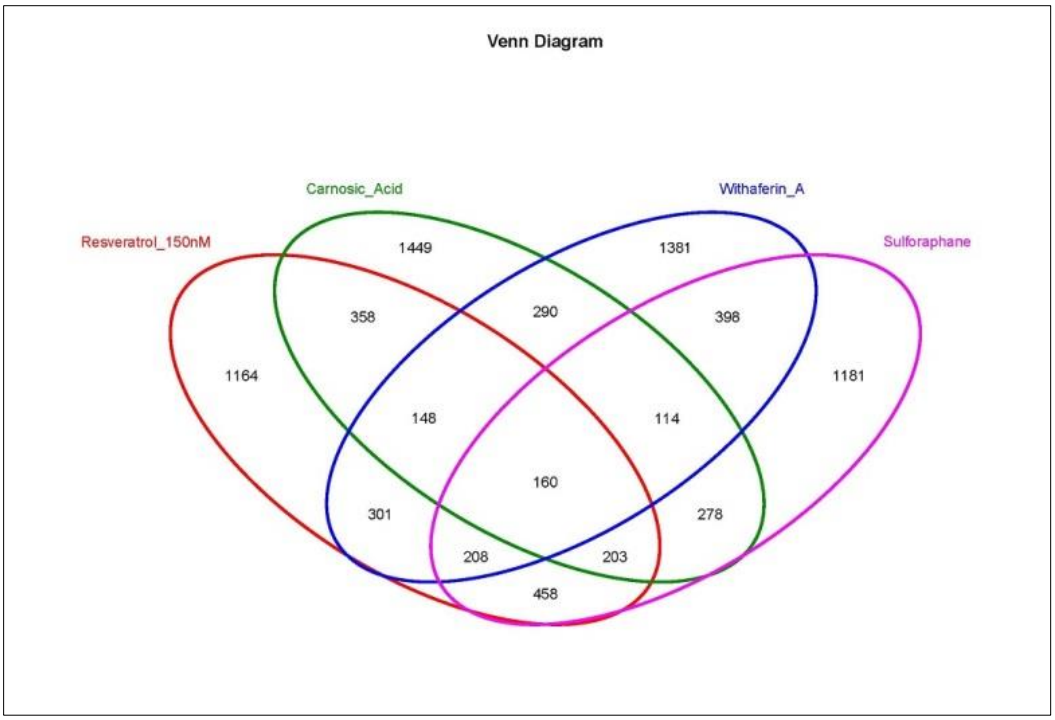


Figura 23: Diagrama de Venn que muestra el solapamiento de los genes diferencialmente expresados en los 4 experimentos seleccionados, tras el tratamiento con distintos compuestos con potenciales propiedades anticancerígenas.

Tabla 7: Resultados del análisis funcional de los 160 genes identificados en común entre 4 experimentos del grupo de compuestos con potenciales propiedades anticancerígenas. La tabla muestra los 20 procesos biológicos más estadísticamente significativos.

Items	Items Details	Support	List size	Reference Support	Reference size	Hyp	Hyp_c	Genes
GO:0000278	mitotic cell cycle (BP)	18	160	304	34208	6.34E-15	4.90E-12	CDC7, KNTC1, PRIM1, TUBA4A, RFC4, NDE1, SKP2, KIF2C, CENPK, CCNB1 , CENPF, GINS2, CENPE, TYMS, CDK4 ,KIF20A , MCM3, NUP107
GO:0051301	cell division (BP)	16	160	286	34208	5.17E-13	2.00E-10	CDC7, KNTC1, HELLS, NDE1, CDCA7L, KIF2C, NCAPD2, KIF11, TIPIN, CCNB1, CENPF, CENPE, TIMELESS, ASPM, CDK4 ,KIF20B
GO:0006260	DNA replication (BP)	12	160	152	34208	8.36E-12	2.15E-09	CDC7, PRIM1, PHB, RFC4, CENPF, GINS2, TYMS, TOP2A, RNASEH2A, RRM1, TNFAIP1, MCM3
GO:0000087	M phase of mitotic cell cycle (BP)	9	160	95	34208	7.12E-10	1.37E-07	KNTC1, NDE1, KIF2C, CENPK, CENPF, CENPE, KIF20A, NUP107, KIF20B
GO:0000236	mitotic prometapha se (BP)	8	160	85	34208	6.81E-09	1.05E-06	KNTC1, NDE1, KIF2C, CENPK, CCNB1, CENPF, CENPE, NUP107
GO:0006916	anti-apoptosis (BP)	9	160	200	34208	4.69E-07	6.04E-05	NFKBIA, HELLS, THBS1, CITED2, HMOX1, PEA15, BIRC3, ADAM17, ANXA4

Items	Items Details	Support	List size	Reference Support	Reference size	Hyp	Hyp_c	Genes
GO:0006979	response to oxidative stress (BP)	7	160	110	34208	9.12E-07	0.000100593	DUSP1, GCLM, TXNRD1, SLC7A11, HMOX1, SRXN1, EPAS1
GO:0007049	cell cycle (BP)	12	160	435	34208	1.09E-06	0.000104992	DUSP1, HELLS, NCAPD2, KIF11, TIPIN, CKAP2, PIM1, TIMELESS, ASPM, E2F8, CDK4, KIF20B
GO:0055086	nucleobase-containing small molecule metabolic process (BP)	6	160	77	34208	1.72E-06	0.000132585	TXNRD1, GLRX, PAICS, TYMS, RRM1, CAD
GO:0042493	response to drug (BP)	10	160	301	34208	1.71E-06	0.00014629	GCLM, THBS1, HMGCS1, CCNB1, CENPF, HADH, TYMS, TOP2A, ADAM17, CDK4
GO:0000723	telomere maintenance (BP)	5	160	50	34208	3.76E-06	0.000241768	PARP1, PRIM1, TINF2, RFC4, XRCC5
GO:0033261	regulation of S phase (BP)	3	160	7	34208	3.47E-06	0.000243295	CDC7, TIPIN, TIMELESS
GO:0007018	microtubule-based movement (BP)	6	160	90	34208	4.29E-06	0.000254751	TUBA4A, KIF2C, KIF11, CENPE, KIF20A, KIF20B
GO:0000082	G1/S transition of mitotic cell cycle	7	160	145	34208	5.77E-06	0.000318248	CDC7, PRIM1, SKP2, CCNB1, TYMS, CDK4, MCM3
GO:0008219	cell death (BP)	7	160	156	34208	9.32E-06	0.000479774	OPTN, TPP1, SLC33A1, HMOX1, TNFRSF12A, CLN3, GBA
GO:0006879	cellular iron ion homeostasis (BP)	5	160	63	34208	1.19E-05	0.000509145	FTH1, ATP6V0D1, HMOX1, MCOLN1, ATP6V0B
GO:0006271	DNA strand elongation involved in DNA replication (BP)	4	160	30	34208	1.15E-05	0.00052171	PRIM1, RFC4, GINS2, MCM3
GO:0007596	blood coagulation (BP)	11	160	457	34208	1.12E-05	0.000539751	MAFG, THBS1, TUBA4A, ITGA3, CABLES1, SLC7A11, KIF2C, KIF11, ITGAV, CENPE, ITGA2
GO:0000084	S phase of mitotic cell cycle (BP)	6	160	112	34208	1.51E-05	0.000615431	PRIM1, RFC4, SKP2, GINS2, CDK4, MCM3
GO:0051987	positive regulation of attachment of spindle microtubules to kinetochore	2	160	2	34208	2.17E-05	0.000839193	CCNB1, CENPE

Gene Symbol	HT29 CarnosicAcid 24h	HT29 RoseMary 24h	MCF10A Sulforaphane 15uM 24h	MCF7 WithaferinA	MDA WithaferinA	MCF7 Resveratrol 150nM	MCF7 Resveratrol 250nM	MDAMB157 I3C	MDAMB231 I3C	MDAMB436 I3C
CCNB1	-0.574825019	-1.66013322	-0.534037482	-1.40324468	-0.8312067	-2.142057825	-2.216753323	-0.570505256	-0.17791673	-0.646853465
CDC7	-0.411155777	-2.01995653	-1.718478125	-1.23459243	-0.91494907	-1.344198042	-2.130631789	0.130304605	-0.02799605	-0.05003585
CDK4	-0.283022151	-0.47564104	-0.703014423	-0.85466943	0.697990477	-0.628077576	-0.58129011	0.064984623	-0.04922452	-0.13314388
CENPE	-0.487766639	-2.00245839	-1.128687876	-0.7224492	-1.18333308	-1.95185863	-2.370018845	-0.297558708	-0.61448207	-0.641839027
CENPF	-0.640244383	-2.1492013	-0.737075892	-0.99456595	-1.04388731	-3.340428647	-3.345271807	-0.447744573	-0.37443161	-0.441948345
CENPK	-0.46739711	-2.16001207	-1.665218433	-1.21919569	-0.71467693	-1.821352536	-2.321863601	-0.229369279	-0.47540822	-0.26628989
GINS2	-0.479186388	-2.33074932	-1.652431996	-1.45237554	-0.96611713	-2.133086981	-2.629570249	-0.078774082	-0.70959553	-0.09656379
KIF20A	-0.63540163	-2.68337503	-0.634316661	-1.62934718	-0.93936548	-3.843728521	-3.466072451	-0.894432684	-0.33351586	0.050724293
KIF2C	-0.373403261	-1.80464038	-0.792116514	-1.13575175	-0.81881777	-2.88792853	-2.8677782	-0.059438466	-0.29750167	-0.050759383
KNTC1	-0.333359258	-1.7874979	-0.711132211	-1.16262096	-1.29945773	-1.601723353	-2.31547688	-0.621220369	-0.36505515	-0.820585949
MCM3	-0.365421123	-1.76536263	-1.268132365	-1.10421698	-1.30107276	-1.773631745	-2.4688302	0.109720073	-0.4102876	-0.481491813
NDE1	-0.313092769	-0.92974773	-0.373919259	-0.11739239	-0.73695146	-1.178599003	-0.809565858	-0.152859317	-0.43225442	-0.539998753
NUP107	-0.535770933	-1.52841848	-0.875240535	-0.79567569	-0.80131932	-0.963164768	-1.655537364	-0.010112352	-0.10240298	-0.179877633
PRIM1	-0.255954527	-1.67095708	-1.393723094	-1.54565623	-0.95546801	-2.281338551	-2.675217941	0.004531427	-0.37721414	-0.290063064
RFC4	-0.324659788	-1.23387771	-0.830988498	-0.95134597	-0.98664665	-1.73933681	-1.670285273	0.15527076	-0.20701296	0.077108473
SKP2	-0.612588936	-2.22539482	-0.846546903	-0.96792151	-0.89909547	-2.086505274	-2.289428345	-0.237805959	-0.23776401	-0.437138767
TUBA4A	0.353973297	0.368228367	0.72475414	0.598128881	0.601194673	0.774883442	0.754442159	0.181099757	0.562820919	0.743967973
TYMS	-0.360309139	-2.66493266	-0.981131722	-1.91556285	-1.90279	-0.836752894	-1.60931121	-0.137377777	-0.52537952	-0.438395463

Tabla 8: Valores de expresión diferencial de los 18 genes implicados en el proceso biológico GO:0000278 ("mitotic cell cycle") en los experimentos incluidos en el grupo de compuestos con potenciales propiedades anticancerígenas. En verde aparecen los valores negativos, que indican represión del gen tras el tratamiento.

Tabla 9: Detalles de los 18 genes incluidos en la firma molecular identificado con propiedades anticancerígenas.

Simbolo de gen	Descripción
CCNB1	Ciclina B1
CDC7	Proteína de ciclo de división celular 7
CDK4	Quinasa 4 dependiente de ciclina
CENPE	Proteína E asociada a centrómero
CENPF	Proteína F asociada a centrómero
CENPK	Proteína K asociada a centrómero
GINS2	Subunidad 2 del complejo GINS
KIF20A	Miembro 20A de la familia de quinesinas
KIF2C	Miembro 2C de la familia de quinesinas
KNTC1	Proteína asociada al quinetocoro 1
MCM3	Componente 3 del complejo de mantenimiento de minicromosoma
NDE1	Proteína 1 de distribución nuclear E
NUP107	Nucleoporina 107
PRIM1	Subunidad 1 de primasa de ADN
RFC4	Subunidad 4 del factor de replicación C
SKP2	Proteína quinasa 2 asociada a la fase S
TUBA4A	Tubulina alfa 4A
TYMS	Sintasa de timidilato

5. LA FIRMA MOLECULAR DE 18 GENES IDENTIFICA COMPUESTOS CON POTENCIAL ANTICANCERÍGENO

Para continuar evaluando el potencial anticancerígeno de la firma molecular identificada, se ha analizado la expresión diferencial de estos genes en todos los experimentos incluidos en la base de datos mediante un mapa de calor (**Figura 24**). Como era de esperar, los compuestos presentes en el grupo identificado como potencialmente anticancerígeno (WFNA, SFN, RVT, and RSM) se agrupan de manera muy próxima en base a la expresión diferencial de estos 18 genes. Los tratamientos con I3C en diferentes células cancerígenas también se agrupan con estos compuestos, pero el nivel de represión de los genes de la firma genómica identificada es variable en función de la línea celular, concretamente es menor en los subtipos de células MDA. El tratamiento con CA en células HT29 se encuentra muy próximo al grupo que engloba los compuestos potencialmente anticancerígenos debido al menor nivel de represión obtenido en los genes de la firma genómica.

Sorprendentemente se observan 2 experimentos no identificados anteriormente dentro del grupo de compuestos potencialmente anticancerígenos: Macrófagos tratados con diferentes cepas de LB rhamnosus y células CaCo2 tratadas con trans-resveratrol (TRVT). Tal observación podría

explicarse por la estrecha relación existente entre los procesos biológicos de ciclo celular y estrés oxidativo, dado que microorganismos como LB han demostrado ejercer efectos positivos sobre el estrés oxidativo celular (85). Del mismo modo, se observa que de los 3 experimentos que emplean un tratamiento con RSM, únicamente el experimento que utiliza la mayor concentración (100µg/mL) se encuentra en el grupo de los compuestos con potencial anticancerígeno. Esta última observación confirma que altas concentraciones de RSM son altamente efectivas para arrestar el ciclo e inducir muerte celular, como ya se ha observado previamente (22).

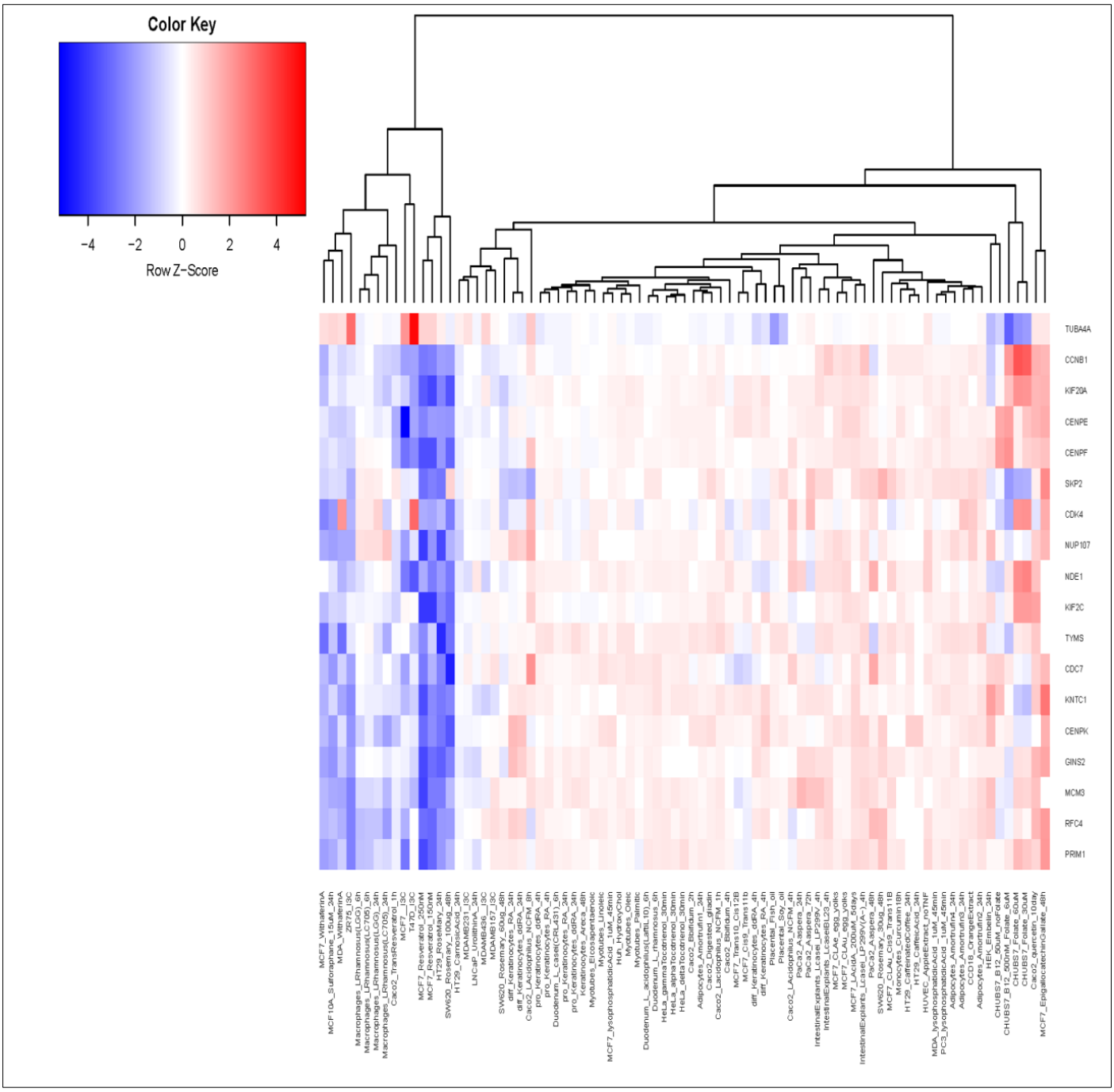


Figura 24: Mapa de calor del nivel de expresión diferencial de los genes incluidos en la firma genómica con potencial anticancerígeno en todos los experimentos de la base de datos. Los valores de log2 FC han sido estandarizados a nivel de las filas.

CAPÍTULO 3: DESARROLLO DE UNA APLICACIÓN WEB PARA MINERÍA DE DATOS EN NUTRIGENÓMICA.

1. VISIÓN GENERAL DE LA APLICACIÓN WEB

La aplicación web de minería de datos en nutrigenómica desarrollada durante este trabajo de doctorado, titulada NutriGenomeDB, se encuentra públicamente disponible y totalmente accesible a través del siguiente enlace: http://nutrigenomedb.org/

NutriGenomeDB está basada en las firmas moleculares obtenidas tras el análisis de los experimentos de nutrigenómica identificados en GEO, y que han cumplido con los criterios de inclusión. De esta manera, la aplicación contiene un total de 61 estudios de nutrigenómica, los cuales representan un total de 231 experimentos de expresión diferencial. Cada uno de estos experimentos se caracteriza por una firma molecular compuesta por el conjunto del 10% de los genes diferencialmente expresados, y ordenados en función de su significancia estadística. De un total de 568.463 genes seleccionados para construir las firmas moleculares albergadas en la aplicación web, 156.374 han obtenido un valor *p* ajustado inferior o igual a 0.05. De este modo, el número de genes incluidos en las firmas moleculares varía entre 1.500 y 3.000, dependiendo del array de expresión utilizado para el análisis experimental.

La aplicación web de minería de datos en nutrigenómica se compone de dos módulos funcionales:

- Módulo exploratorio: permite explorar el nivel de expresión diferencial de genes a lo largo de todos los experimentos incluidos en la base de datos. Los resultados se presentan en forma de tablas y visualizaciones interactivas, las cuales pueden ser descargadas por el usuario.

- Módulo analítico: permite comparar y cuantificar la similitud entre una firma molecular externa con las firmas moleculares obtenidas de los experimentos de nutrigenómica presentes en la base de datos. Su objetivo es conectar los mecanismos moleculares que provocan tratamientos específicos con drogas o medicamentos, con aquellos provocados por los compuestos con potencial uso en alimentación. Las conexiones identificadas se pueden cuantificar mediante el número de genes en común entre las 2 firmas moleculares, y también mediante la puntuación NES. Un valor NES elevado revela que los genes identificados en común se encuentran fuertemente sobreexpresados en ambas firmas moleculares, y viceversa.

2. EXPLORACIÓN DE LOS NIVELES DE EXPRESIÓN DIFERENCIAL: EJEMPLO DE LA FIRMA MOLECULAR POTENCIALMENTE ANTICANCERÍGENA IDENTIFICADA

El módulo exploratorio puede ser interrogado utilizando uno o varios símbolos de genes humanos. Dispone de una consulta de ejemplo, y de carga automática, que incluye un conjunto de genes implicados en el metabolismo del colesterol. El resultado presenta una tabla interactiva que incluye únicamente los experimentos donde los genes interrogados aparecen incluidos en su respectiva firma molecular. Se muestra la correspondiente información sobre los niveles de expresión diferencial obtenidos (valores de log2 FC), información estadística del análisis de los datos, así como información adicional sobre el diseño experimental y el estudio de origen (compuesto utilizado, concentraciones si estuvieran disponibles, identificador del estudio original en GEO). Cada experimento de la tabla obtenida se encuentra directamente enlazado al estudio que lo ha generado, mediante el identificador de GEO correspondiente. Para la inspección visual de estos resultados, es posible generar un gráfico de líneas, y descargarlo localmente en formato PNG (**Figura 25**). Para los genes interrogados no incluidos en determinadas firmas moleculares, el gráfico muestra un valor de expresión diferencial log2 FC igual a 0. De esta manera es posible identificar visualmente patrones de co-expresión entre los genes interrogados.

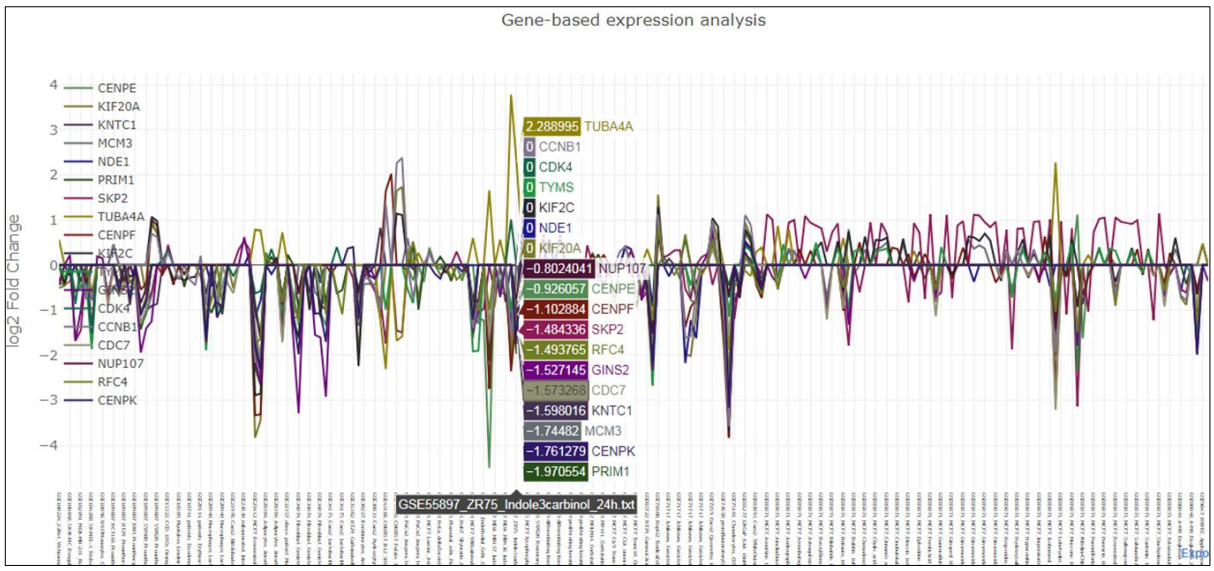


Figura 25: Captura de pantalla del gráfico de líneas generado en el módulo exploratorio de NutriGenomeDB. La consulta está compuesta por los 18 genes identificados en la firma molecular potencialmente anticancerígena identificada en el punto 5 del Capítulo 2. El gráfico generado es completamente interactivo y permite identificar patrones de co-expresión entre varios genes. En este ejemplo, el tratamiento de células ZR-75 con I3C durante 24h (GSE55897) provoca la represión simultánea de 11 de los 18 genes interrogados.

3. ANÁLISIS DE FIRMAS MOLECULARES EXTERNAS

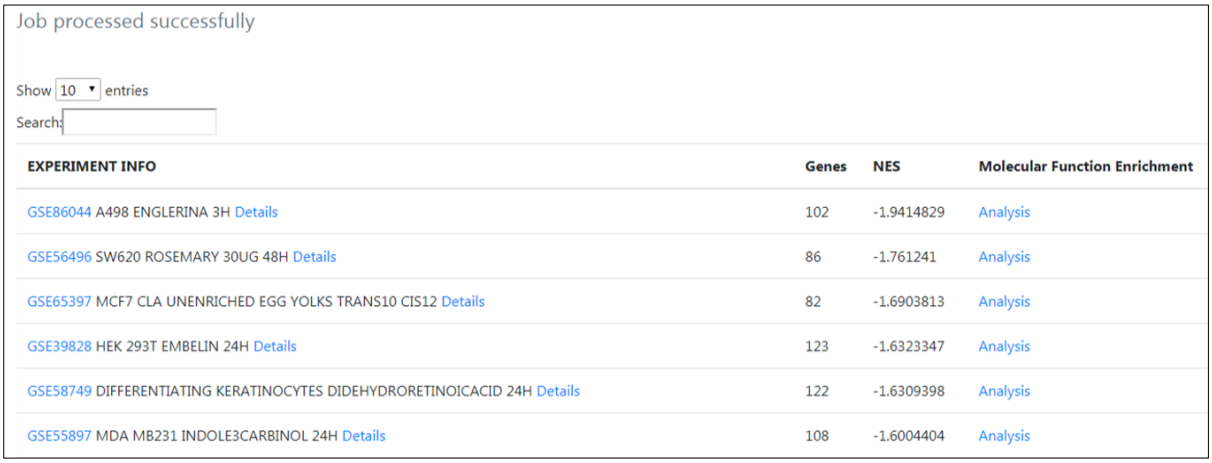
El módulo analítico permite identificar potenciales conexiones entre los experimentos de nutrigenómica incluidos en la base de datos, y un perfil de expresión o firma molecular externa introducido en el módulo, característico de algún fenotipo determinado. La esencia de esta aproximación es que los alimentos y sus compuestos bioactivos podrían ejercer sus propiedades saludables actuando en las mismas dianas moleculares, o mediante los mismos mecanismos moleculares, que las drogas o medicamentos utilizados para el tratamiento de determinadas enfermedades.

Para lanzar una consulta, los datos introducidos deben contener dos columnas: símbolos de genes humanos y valores de expresión diferencial en escala logarítmica en base 2. La aplicación ofrece unos datos de prueba de carga automática para ilustrar el formato requerido, correspondientes a un experimento tras un tratamiento con Metformina, un fármaco utilizado para el tratamiento de la diabetes de tipo 2. Los datos introducidos son comparados con las firmas moleculares de NutrigenomeDB utilizando el algoritmo GSEA de comparación de patrones de firmas moleculares (68). Los resultados obtenidos se presentan en una tabla interactiva, la cual informa sobre los experimentos de nutrigenómica de la base de datos potencialmente conectados con los datos externos introducidos, el número de genes identificados en común, así como información sobre el nivel de enriquecimiento en los genes sobreexpresados o reprimidos (puntuación NES). La tabla interactiva puede ser ordenada por el usuario en función de cualquier columna. Desde la misma tabla de resultados, también es posible lanzar un análisis de enriquecimiento de las funciones moleculares sobre representadas de los genes en común, con el fin de identificar los mecanismos moleculares compartidos entre la firma molecular externa y la firma molecular del experimento de nutrigenómica potencialmente conectado.

4. CASO DE USO: EL FÁRMACO AMLODIPINO

Para ilustrar la utilidad del módulo analítico de NutriGenomeDB, se han elegido unos datos de expresión génica obtenidos tras un tratamiento con Amlodipino, un fármaco comúnmente utilizado para el tratamiento de la hipertensión. Los datos provienen de un experimento de la base de datos GEO, con identificador GSE42808, que estudia el efecto del Amlodipino en la expresión génica de células humanas de endotelio de cordón umbilical. A partir de la lista de genes diferencialmente expresados en este experimento, se han seleccionado los 1.000 genes con mayor significancia estadística para definir la firma molecular que provoca el tratamiento con Amlodipino en las células utilizadas.

Tras lanzar la consulta desde NutriGenomeDB con los 1.000 genes identificados, la tabla de resultados informa sobre potenciales conexiones entre la firma molecular del Amlodipino y experimentos de nutrigenómica que emplean compuestos tales como la Englerina A, un extracto de RSM o el ácido linoleico conjugado trans-10,cis-12 (**Figura 26**). La tabla de resultados ha sido ordenada por orden creciente de la puntuación NES, con lo cual las conexiones del principio de la tabla informan sobre experimentos de nutrigenómica que provocan una fuerte represión de genes similar al tratamiento con el fármaco Amlodipino.


Job processed successfully

Show 10 entries

Search:

EXPERIMENT INFO	Genes	NES	Molecular Function Enrichment
GSE86044 A498 ENGLERINA 3H Details	102	-1.9414829	Analysis
GSE56496 SW620 ROSEMARY 30UG 48H Details	86	-1.761241	Analysis
GSE65397 MCF7 CLA UNENRICHED EGG YOLKS TRANS10 CIS12 Details	82	-1.6903813	Analysis
GSE39828 HEK 293T EMBELIN 24H Details	123	-1.6323347	Analysis
GSE58749 DIFFERENTIATING KERATINOCYTES DIDEHYDRORETINOICACID 24H Details	122	-1.6309398	Analysis
GSE55897 MDA MB231 INDOLE3CARBINOL 24H Details	108	-1.6004404	Analysis

Figura 26: Resultados obtenidos en el módulo analítico de NutriGenomeDB. La tabla presenta una lista de experimentos de nutrigenómica potencialmente conectados con la firma molecular externa introducida. Para cada experimento se muestra el número de genes en común (Genes), el valor de la puntuación NES (NES) e información experimental directamente enlazada al estudio de referencia. Los niveles de expresión de los genes identificados en común pueden ser inspeccionados a partir del enlace "Details", situado a la derecha de la información experimental. El enlace de "Analysis", situado en la columna de la derecha, permite lanzar un análisis de enriquecimiento funcional de funciones moleculares a partir de los genes identificados en común.

Con el fin de identificar los mecanismos moleculares que explicarían las conexiones identificadas, desde la interfaz web de NutriGenomeDB se han lanzado distintos análisis de enriquecimiento funcional de los genes en común. El primer análisis, utilizando los 102 genes en común con el experimento que utiliza el producto natural Englerina A ("GSE86044 A498 Englerin A 3h"), no ha revelado ninguna función molecular sobre-representada estadísticamente significativa. Sin embargo el segundo análisis funcional, utilizando los 86 genes en común identificados en el segundo experimento realizado con RSM ("GSE56496 SW620 Rosemary 30µg 48h") ha revelado información potencialmente interesante (**Figura 27**). Las funciones moleculares más enriquecidas, y estadísticamente significativas, están estrechamente relacionadas con actividades de transporte de iones a través de las membranas celulares, y particularmente iones de Calcio. Efectivamente estos resultados encajan con el mecanismo molecular del fármaco Amlodipino, el cual actúa como bloqueador de los canales de Calcio para ejercer sus propiedades hipotensoras.

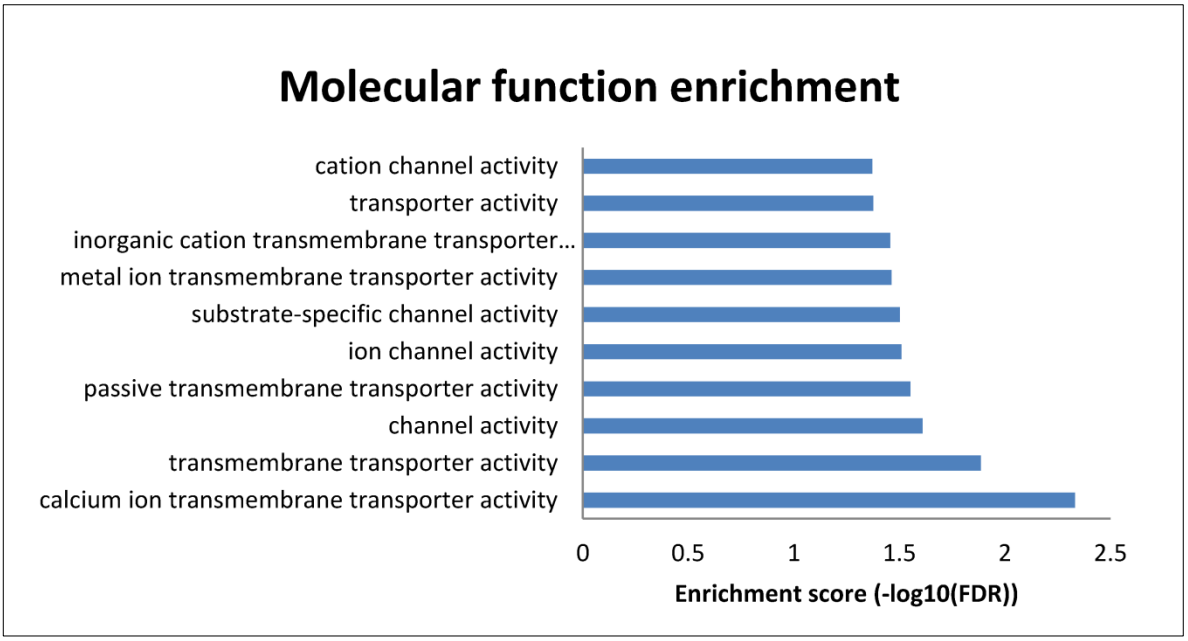


Figura 27: Representación en gráfico de barras de los resultados obtenidos desde NutriGenomeDB tras el enriquecimiento funcional de funciones moleculares. Se han utilizado los 86 genes en común identificados entre la firma molecular provocada por el tratamiento con Amlodipino en células humanas de endotelio de cordón umbilical, y la obtenida en el experimento GSE56496 tras un tratamiento con RSM en células SW620. Sólo se muestran las funciones moleculares enriquecidas con significancia estadística (FDR <= 0.05).

A partir del enlace "Details" del módulo analítico de NutriGenomeDB, también es posible inspeccionar el perfil de expresión diferencial de los genes en común entre la firma molecular interrogada y los experimentos de nutrigenómica potencialmente conectados. Como era de esperar, entre los genes relacionados con funciones moleculares de transporte transmembrana, algunos se encuentran fuertemente reprimidos en ambas firmas moleculares (**Tabla 10**).

Tabla 10: Genes con funciones moleculares relacionadas con el transporte transmembrana de iones Calcio, y comúnmente reprimidos entre la firma molecular provocada por el tratamiento con Amlodipino en células humanas de endotelio de cordón umbilical, y el experimento de nutrigenómica altamente conectado correspondiente a un tratamiento con RSM en células SW620 (GSE56496).

Gene Symbol	log2FC Amlodipine	log2FC GSE56496 Rosemary 30ug 48h	Description
TRPV2	-1.327098208	-0.674827345	Receptor 2 transitorio del canal de cationes subfamilia V miembro 2
ATP2C2	-1.413693464	-1.119488698	ATPasa secretora de la vía metabólica Ca2+ transportina 2
GRIK2	-1.329016735	-1.027980433	Subunidad 2 de tipo kainato del receptor ionotrópico de glutamato
SORCS3	-2.566320697	-3.12378339	Receptor 3 con dominio VPS10 relacionado con sortilina

5. AGRUPAMIENTO DE COMPUESTOS PRESENTES EN LA BASE DE DATOS MEDIANTE MAPAS DE CALOR

NutriGenomeDB permite generar mapas de calor desde su módulo exploratorio, lanzando una consulta a partir de varios símbolos de genes. Los gráficos obtenidos son completamente interactivos y permiten inspeccionar detalladamente cada uno de los clústeres y patrones de expresión identificados.

A partir de los resultados obtenidos tras el análisis de enriquecimiento funcional de los genes en común identificados entre las firmas moleculares provocadas por el fármaco Amlodipino y el experimento "GSE56496 SW620 Rosemary 30µg 48h", se han seleccionado 11 genes anotados con una función molecular relacionada con el transporte transmembrana. Estos genes han sido utilizados para generar un mapa de calor, con el objetivo de encontrar grupos de compuestos con potencial uso en alimentación capaces de modular su expresión. El mapa de calor obtenido permite identificar un grupo formado por 16 experimentos de nutrigenómica, en los cuales al menos uno de los genes interrogados, relacionados con el transporte transmembrana, se encuentra incluido en la firma molecular que lo define (**Figura 28**).

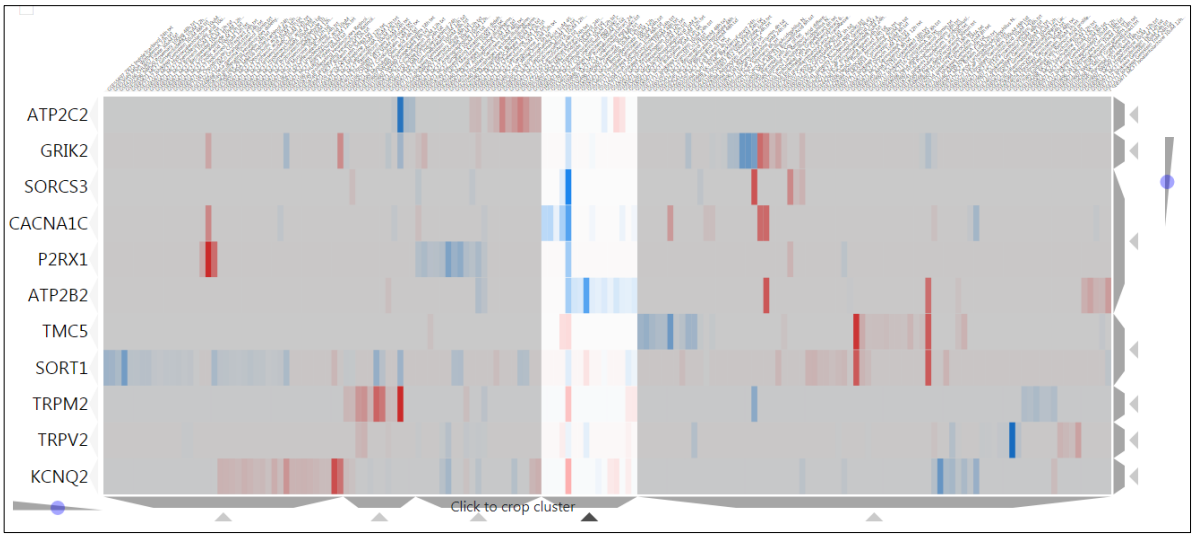


Figura 28: Visión general del mapa de calor generado desde NutriGenomeDB a partir de una selección de 11 genes implicados en el transporte transmembrana. El color azul representa represión en la expresión diferencial, y rojo representa la sobreexpresión. El gráfico es completamente interactivo y permite seleccionar los distintos grupos identificados para su inspección detallada. En la imagen se ha seleccionado un grupo de 16 experimentos de nutrigenómica que presenta interesantes patrones de represión.

La inspección detallada del grupo seleccionado (**Figura 29**) muestra que el gen ATP2B2 ("ATPase secretory pathway Ca2+ transporting 2") presenta una represión de su expresión consistente en 11 de los 16 experimentos de nutrigenómica agrupados. En el experimento GSE34074, correspondiente a un tratamiento de Fibroblastos con 100µM de genisteína durante 24h, se observa una importante represión de este gen. Esta observación es interesante ya que previamente se ha demostrado, utilizando distintos modelos animales, el potencial de la genisteína como compuesto para el tratamiento de la hipertensión (86-88). Paradójicamente, el mismo gen ATP2B2 también ha sido identificado en la firma molecular del fármaco Amlodipino, pero se encuentra ligeramente sobreexpresado.

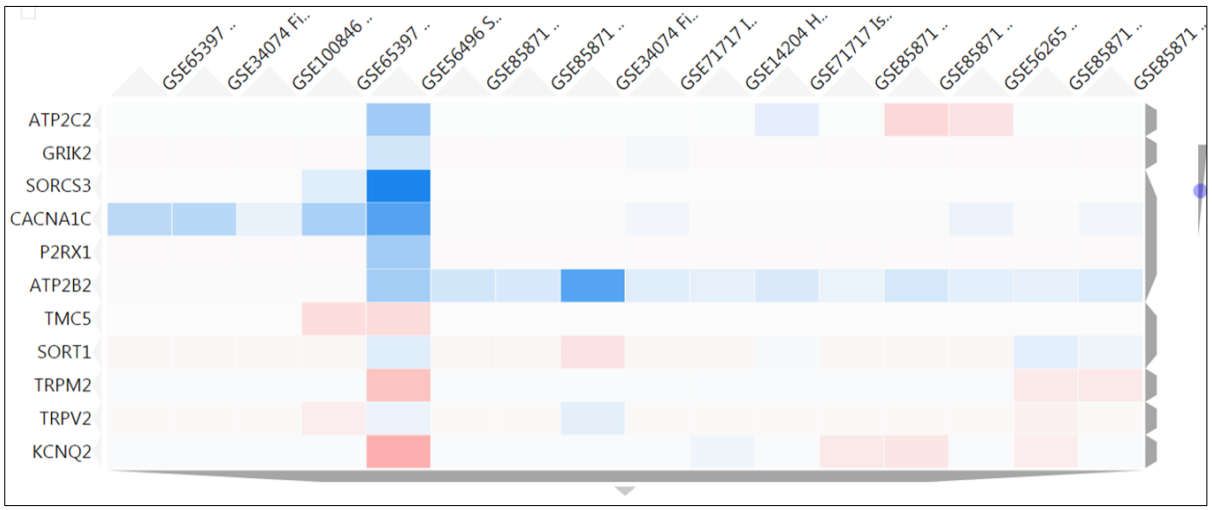


Figura 29: Visión detallada del grupo de 16 experimentos de nutrigenómica identificados en el mapa de calor, generado a partir de una selección de 11 genes con implicaciones en el transporte transmembrana. El experimento GSE56496, a partir del cual se han seleccionado los 11 genes, contiene valores de expresión diferencial para cada uno de ellos. El gen ATP2B2 muestra una represión consistente en 11 de los experimentos incluidos en el grupo seleccionado. GSE34074 (tratamiento de Fibroblastos con 100µM de Genisteína durante 24h) presenta la represión más importante de este gen.

DISCUSIÓN

Las intervenciones nutricionales han demostrado la capacidad de los alimentos y sus compuestos bioactivos para regular la expresión de los genes (89,90), y su potencial efecto protector para la prevención de determinadas enfermedades (91-93). Los datos de transcriptómica que se acumulan en los repositorios públicos, dónde cada vez se encuentran disponibles más experimentos de nutrigenómica, presentan un amplio potencial para profundizar en el conocimiento de los mecanismos moleculares que influyen en la capacidad de la dieta para modular la salud. Este trabajo doctoral se ha centrado en el análisis de experimentos de nutrigenómica, recopilando datos de expresión génica disponibles en la base de datos GEO con el fin de realizar un análisis integrativo que permitiera extraer nuevos conocimientos. Todo ello se traduce en la publicación en línea de una plataforma web de minería de datos en nutrigenómica, disponible públicamente, que permite explorar el potencial nutrigenómico de los compuestos incluidos, así como conectar firmas moleculares externas con los experimentos de nutrigenómica. A partir de la comparación entre los perfiles de expresión génica obtenidos en experimentos de nutrigenómica y aquellos provocados por fármacos utilizados para el tratamiento de enfermedades, la plataforma de minería de datos desarrollada permite asignar potenciales efectos beneficiosos a alimentos y compuestos bioactivos. Además, gracias al análisis funcional de los genes coincidentes entre 2 perfiles de expresión diferencial o firmas moleculares, es posible aportar una explicación molecular para las propiedades saludables asignadas, mediante la identificación de potenciales mecanismos moleculares responsables.

EL HIDROXITIROSOL MODULA LA EXPRESIÓN DE MICRO ARN'S EN EL HÍGADO DE RATONES

Si bien cada vez más estudios evidencian que los compuestos bioactivos de los alimentos son capaces de modular la expresión de miARN's en modelos animales (94,95), y en humanos (96), los estudios que han analizado el efecto del HT en la expresión de los elementos que componen el genoma son muy escasos. Por ejemplo, se ha estudiado el efecto del HT sobre miARN's específicos *in vitro*: los miR-9 (97) y miR-146a (98). Hasta ahora, únicamente un estudio ha evaluado el efecto del HT en el miRNoma completo del intestino delgado de ratón (99).

De entre los miARN's identificados como diferencialmente expresados en el estudio presentado en este trabajo (**Tabla 5**), en respuesta al consumo de una dieta suplementada con HT, el miR-802-5p ha sido previamente descrito como un actor importante en la desregulación del metabolismo de la glucosa, participa en la señalización celular por angiotensina, y se encuentra sobreexpresado en sujetos obesos (100,101). Curiosamente, en un estudio previo, este miRNA ha

mostrado una regulación similar, esta vez en intestino de ratones (99). La regulación de dicho miRNA en dos tejidos diferentes indica que este podría ser una diana potencial del hidroxitirosol. En el caso de miR-423-3p, sus niveles de expresión han sido asociados positivamente con el crecimiento celular en casos de cáncer de hígado y colon (102,103). En cuanto al miR-30a-5p, se ha descrito que su inducción mejora procesos de fibrosis en el hígado (104), y también suprime el crecimiento de cáncer mamario y metástasis (104,105). Finalmente el miR-146b ha demostrado su potencial para disminuir la esteatosis hepática no-alcohólica (106), aunque también su represión podría promover el crecimiento de células cancerígenas y la metástasis (107). El análisis de enriquecimiento funcional realizado a partir de los genes diana de estos 4 miARN's (**Figura 19**), utilizando anotaciones de vías metabólicas de la base de datos Panther, no ha hecho más que confirmar su potencial implicación en la regulación de importantes vías de señalización celular: señalización Wnt (P00057), señalización por el receptor de colecistoquinina B (CCKR) (P06959) y señalización de inflamación a través de citoquinas y quimiocinas (P00031). Estos resultados se reflejan en el segundo trabajo de investigación publicado presente en el apartado **Anexo**.

El análisis funcional de los miARN's no es un proceso estandarizado, y por ello los resultados obtenidos pueden variar considerablemente en función del método elegido. El análisis funcional de estas moléculas se realiza a partir de sus genes diana. Para la búsqueda de genes diana, existen multitud de algoritmos que utilizan diferentes esquemas de puntuación para realizar predicciones de interacción entre un miARN y su gen diana. La coincidencia por complementariedad de secuencia entre un transcrito y la región semilla del miARN es el parámetro que generalmente aporta mayor peso a la puntuación obtenida. Además algunos algoritmos también incorporan parámetros con información sobre la estabilidad energética de la interacción identificada, o el nivel de conservación de las secuencias semillas de los miARN's (108,109). Incluso se ha desarrollado un algoritmo de aprendizaje automático para cuantificar la probabilidad de que se produzca una interacción entre el miARN y su ARNm diana, a partir de un modelo previamente entrenado con distintas características de las secuencias de interacción predichas por dicho algoritmo (110). Además existe una base de datos que alberga información sobre interacciones validadas experimentalmente por distintas técnicas de biología molecular, aunque la cantidad de genes diana validados experimentalmente suele ser escasa e insuficiente para su análisis funcional en una etapa posterior (111). También se ha observado que la interacción *in vivo* puede producirse en distintas posiciones del ARNm diana, aumentando así la complejidad para obtener predicciones fiables y que puedan producirse *in vivo* (112). Generalmente los resultados de los distintos algoritmos de predicción devuelven una cantidad importante de genes diana, y estos suelen ser filtrados en función de las coincidencias encontradas entre distintos algoritmos.

El reducido número de miARN's diferencialmente expresados obtenidos tras el experimento podría explicarse por distintos aspectos de la biogénesis de estas moléculas así como por la técnica de detección empleada. Por ejemplo se ha descrito que, dependiendo del tejido analizado, muchos miARN's son transcritos pero no son procesados para obtener su forma madura (113). En el estudio presentado en este trabajo, el análisis de los datos obtenidos por secuenciación se ha centrado en la detección de miARN's en su forma madura, ya que es de esta manera como estas moléculas ejercen su potencial efecto biológico. Sin embargo, son necesarios más estudios en distintos modelos sobre el potencial efecto biológico provocado por la capacidad del HT para regular la expresión de los miARN's en el hígado y otros tejidos.

IDENTIFICACIÓN DE COMPUESTOS POTENCIALMENTE ANTICANCERÍGENOS POR AGRUPAMIENTO JERÁRQUICO Y OBTENCIÓN DE UNA FIRMA MOLECULAR CON PROPIEDADES ANTICANCERÍGENAS

La búsqueda de datos de experimentos de nutrigenómica en la base de datos GEO ha permitido construir la primera versión de una base de datos que contiene los niveles de expresión diferencial de genes en respuesta al tratamiento con distintos alimentos y sus compuestos bioactivos. Con el objetivo de conectar los distintos tratamientos a través de la identificación de potenciales mecanismos moleculares en común, se ha realizado un análisis integrativo de estos datos, utilizando la técnica de agrupamiento jerárquico. Los resultados han mostrado que el componente histológico es el principal factor dominante en el proceso de agrupamiento (**Figura 21**).

Sin embargo hemos sido capaces de identificar un grupo de experimentos, realizados en distintas células de origen cancerígeno y tratadas con compuestos alimenticios que previamente han demostrado poseer propiedades anticancerígenas. El análisis independiente de los 4 experimentos más representativos de este grupo ha revelado un importante solapamiento entre los 3.000 genes diferencialmente expresados y ordenados por significancia estadística, obteniendo así en diagrama de Venn con 160 genes en común (**Figura 23**). El análisis de enriquecimiento funcional de estos genes ha desvelado el proceso biológico de ciclo celular mitótico (GO:0000278) como el más sobre-representado y estadísticamente significativo (**Tabla 7**). Efectivamente este proceso biológico se encuentra estrechamente relacionado con procesos cancerígenos, los cuales se caracterizan por una división celular descontrolada. Los 18 genes anotados en este proceso biológico (**Tabla 9**), y que se encuentran fuertemente reprimidos en los experimentos de nutrigenómica identificados, podrían representar una firma molecular con propiedades potencialmente anticancerígenas.

LA FIRMA MOLECULAR IDENTIFICADA AGRUPA ALIMENTOS Y COMPUESTOS BIOACTIVOS CON PROPIEDADES ANTICANCERÍGENAS PREVIAMENTE DESCRITAS

En los experimentos incluidos dentro del grupo de compuestos con potenciales propiedades anticancerígenas, la expresión diferencial de los 18 genes de la firma molecular identificada muestra un claro patrón de represión (**Tabla 8**). Los compuestos que provocan esta firma molecular se encuentran naturalmente presentes en el broccoli (SFN) y plantas de distintas familias (I3C, WFNA, CA and RVT). En efecto, la planta medicinal Withania somnifera, principal fuente de WFNA, se ha usado tradicionalmente en la medicina ayurvédica.

Para evaluar el potencial de la firma molecular identificada, se ha generado un mapa de calor a partir de los valores de expresión de los 18 genes de la firma molecular identificada (**Figura 24**). El resultado obtenido demuestra que esta firma molecular permite reagrupar juntos los compuestos identificados previamente en el grupo con propiedades anticancerígenas mediante el agrupamiento jerárquico. Este resultado es significativo, ya que los experimentos incluidos en el grupo con propiedades anticancerígenas (sin tener en cuenta los experimentos correspondientes al tratamiento con RVT) identificado en el agrupamiento jerárquico han sido generados en laboratorios distintos, y utilizando 4 arrays de expresión diferentes (GPL10904, GPL16686, GPL10558, GPL4133). Esto demuestra que ha sido posible detectar una relación biológica estable entre los compuestos incluidos en el grupo con propiedades anticancerígenas, salvando así los potenciales efectos-lote que pudieran estar afectando el análisis de agrupamiento jerárquico.

Los 18 genes identificados en la firma molecular se encuentran mayoritariamente reprimidos tras los tratamientos. Estos genes están implicados en importantes procesos biológicos tales como la mitosis y el control del ciclo celular. Codifican proteínas requeridas para la replicación del ADN, el alineamiento y segregación de los cromosomas durante la mitosis (CENPE, CENPF, CENPK, GINS2, MCM3, KNTC1, PRIM1, RFC4, TYMS), el ensamblaje y organización de microtúbulos (NDE1, TUBA4A, KIF2C), proteínas que participan en la señalización celular a través de eventos de fosforilación (SKP2, CDK4, CDC7), una ciclina (CCNB1) y una nucleoporina (NUP107). Además resulta particularmente interesante señalar la represión del gen KIF20A, perteneciente a la familia de las kinesinas y cuyos altos niveles de expresión han sido recientemente asociados a un mal pronóstico en la enfermedad de cáncer (114). Estos resultados se reflejan en el primer trabajo de investigación publicado presente en el apartado **Anexo**.

NUTRIGENOMEDB: CONECTANDO FÁRMACOS Y COMPUESTOS ALIMENTICIOS

Desde la presentación del principio y metodología del CMap, distintos trabajos han demostrado su utilidad, principalmente en el campo del descubrimiento de drogas y reposicionamiento de medicamentos (67,115). Sin embargo, esta metodología nunca había sido aplicada al estudio del potencial nutrigenómico de los alimentos y sus compuestos bioactivos, con el fin de elucidar los mecanismos moleculares que gobiernan el binomio dieta-salud. Con este objetivo, en este trabajo de doctorado se ha desarrollado una aplicación web para minería de datos en nutrigenómica, disponible públicamente en el enlace www.nutrigenomedb.org para su libre utilización por la comunidad científica.

Las firmas moleculares incluidas en la aplicación NutriGenomeDB han sido obtenidas tras un proceso de análisis y tratamiento de datos cuidadoso, sobre todo a nivel del diseño experimental. Los experimentos de nutrigenómica incluidos cumplen con criterios específicos, y los resultados de expresión diferencial no han sido filtrados en base a criterios como la significancia estadística. Efectivamente existen aplicaciones capaces de extraer firmas moleculares de manera automática a partir de la base de datos de GEO, aunque han demostrado presentar importantes sesgos, debido a la asignación automática que realizan de las distintas muestras a los grupos control y tratamiento (64,65). Este aspecto podría resolverse armonizando la anotación de los experimentos disponibles en GEO, cuyos requerimientos de metadatos son actualmente mínimos. La base de datos de firmas moleculares de experimentos de nutrigenómica generada en este trabajo, curada manualmente y que alimenta la plataforma de minería de datos NutriGenomeDB, representa un valioso recurso para la comunidad científica. Presenta un gran potencial para encontrar conexiones entre alimentos, sus compuestos bioactivos y medicamentos empleados para el tratamiento de determinadas enfermedades en base al perfil de expresión diferencial de genes. De esta manera, NutriGenomeDB permite confirmar y formular nuevas hipótesis de investigación sobre los potenciales efectos saludables de determinados alimentos y compuestos bioactivos, además de ofrecer la posibilidad de aportar una explicación molecular para explicar las conexiones identificadas.

Tras introducir en el módulo analítico de NutriGenomeDB la firma molecular provocada por el fármaco Amlodipino, utilizado para el tratamiento de la hipertensión, se ha identificado una conexión potencialmente interesante con un experimento tras el tratamiento de células humanas SW620 con un extracto de RSM (experimento GSE56496) (**Figura 26**). Esta conexión presenta una alta puntuación NES negativa, informando así de coincidencias entre los genes más reprimidos entre ambos experimentos. Este resultado tiene sentido desde un punto de vista biológico, ya que

efectivamente existen numerosas referencias en la literatura científica sobre los potenciales beneficiosos del RSM para el tratamiento de la hipertensión (116-118).

NUTRIGENOMEDB: IDENTIFICANDO MECANISMOS MOLECULARES COMUNES Y AGRUPANDO COMPUESTOS

Desde la misma interfaz de usuario de NutriGenomeDB, es posible realizar un análisis de enriquecimiento funcional de los genes que conectan una firma molecular externa y los experimentos de nutrigenómica incluidos en la base de datos. Este análisis se centra en encontrar las funciones moleculares sobre representadas en los genes diferencialmente expresados en común, con el fin de identificar potenciales mecanismos moleculares que puedan explicar las conexiones obtenidas desde un sentido biológico.

Las funciones moleculares que explicarían la conexión identificada entre el Amlodipino y el experimento realizado en células humanas SW620 con un extracto de RSM están relacionadas con actividades de transporte de iones de Calcio a través de las membranas celulares (**Figura 27**). Efectivamente este mecanismo encaja con el mecanismo de acción del fármaco Amlodipino, el cual es un bloqueador de los canales de Calcio.

A partir de los genes en común entre ambas firmas, y específicamente aquellos genes responsables de actividades de transporte transmembrana identificados tras el enriquecimiento funcional, se ha generado un mapa de calor desde la interfaz del módulo exploratorio de NutriGenomeDB (**Figuras 28 y 29**). De esta manera se han identificado otros experimentos de nutrigenómica en la base de datos generada, capaces de reprimir los genes introducidos, y que por consiguiente también presentarían potenciales propiedades hipotensoras, similares a las demostradas por el RSM. Concretamente, de esta manera se ha conseguido identificar a la genisteína como un compuesto potencial para el tratamiento de la hipertensión. Efectivamente este fitoestrógeno ya ha demostrado previamente sus beneficios para el tratamiento de la hipertensión (119).

De este modo podemos confirmar que la plataforma de minería de datos NutriGenomeDB es capaz de identificar potenciales relaciones biológicas entre los alimentos y compuestos bioactivos incluidos en la base de datos, más allá de las relaciones que pudieran existir a nivel del diseño experimental de distintos experimentos. Así pues, NutriGenomeDB representa un valioso recurso para la comunidad científica y permite generar nuevas hipótesis de investigación para el estudio de la nutrición de precisión. Puede abrir la puerta a mejorar las formulaciones de nuevos alimentos

funcionales con propiedades protectoras frente a determinadas enfermedades, y además permite profundizar en el estudio de los mecanismos moleculares responsables de dichas propiedades.

El efecto biológico de los alimentos y sus compuestos bioactivos depende en gran medida de los procesos fisiológicos de absorción, transporte e interacción con determinados receptores celulares. Sin embargo la identificación de los componentes de los alimentos que presentan propiedades potencialmente protectoras es un paso fundamental para realizar estudios de biodisponibilidad, y poder beneficiarse de sus potenciales propiedades saludables. Estos resultados se reflejan en el tercer trabajo de investigación publicado presente en el apartado **Anexo**.

LIMITACIONES Y PERSPECTIVAS FUTURAS DE NUTRIGENOMEDB

El procesamiento de las firmas moleculares albergadas por la base de datos de NutriGenomeDB es una etapa crítica para el cálculo de las puntuaciones ES. Debido a la diversidad de los arrays de expresión analizados, el tamaño de las firmas moleculares presentes en NutriGenomeDB no es idéntico, y está comprendido entre 1.500 y 3.000 genes. Este aspecto ha podido ser controlado utilizando el valor NES, que tiene en cuenta los tamaños de las firmas moleculares presentes en la base de datos, a la hora de cuantificar las conexiones identificadas.

Del mismo modo, las firmas moleculares obtenidas, disponibles en un archivo en formato GMT, han sido ordenadas de manera decreciente en función del nivel de expresión diferencial de los genes incluidos, antes de eliminar el dato de log2 FC. Por defecto, durante el proceso de comparación de firmas, el algoritmo GSEA divide en 2 cada firma molecular de la base de datos, asumiendo que los genes de la primera mitad han sido sobreexpresados, y los genes de la segunda mitad han sido reprimidos. Durante la definición de las firmas moleculares incluidas en la plataforma de minería de datos en nutrigenómica, no se ha restringido la lista de genes para que la proporción de sobreexpresados y reprimidos fuera exactamente igual, aunque tampoco se han detectado importantes desequilibrios en este sentido. Sin embargo este aspecto tiene una influencia pequeña a la hora de calcular la puntuación NES, ya que los genes que más contribuyen a este valor, según el sistema de puntuación ponderado utilizado en NutriGenomeDB con el algoritmo GSEA, son los que se encuentran en ambos extremos de la lista que representa las firmas moleculares.

El módulo analítico de NutriGenomeDB permite cuantificar las conexiones identificadas de dos maneras: el número de genes diferencialmente expresados en común entre un perfil de expresión diferencial introducido y las firmas moleculares que caracterizan los experimentos de nutrigenómica de la base de datos, y el valor de la puntuación NES. Sin embargo no devuelve ningún valor de significancia estadística sobre las conexiones encontradas. Originalmente el algoritmo GSEA

realiza un análisis de expresión diferencial utilizando los datos de expresión de todas las muestras incluidas en un experimento, los cuales incluyen tantas columnas como muestras analizadas. El programa GSEA cuenta con una variedad de métodos implementados para realizar el análisis de expresión diferencial (estadístico *t*, proporción de clases, diferencia de clases). Posteriormente se cuantifican las conexiones con una base de datos de firmas moleculares a elegir y disponible en línea. Tras ello la significancia estadística de las conexiones encontradas se calcula realizando permutaciones de las muestras incluidas en el experimento analizado, asignándolas aleatoriamente al grupo control o tratamiento. Sin embargo la implementación del algoritmo GSEA en la plataforma de minería de datos desarrollada en este trabajo utiliza el modo "pre-ranked" como parámetro. De esta manera se evita la etapa de cálculo de expresión diferencial por el algoritmo GSEA, muy criticada por ciertos autores (120). Utilizando el modo "pre-ranked", el cálculo de la significancia estadística de las conexiones encontradas se realiza mediante permutaciones de los genes incluidos en las firmas moleculares de la base datos. Según el equipo desarrollador de GSEA, este método estaría sesgado, ya que la creación de firmas moleculares aleatorias perturba las potenciales correlaciones entre genes, con lo cual el valor de significancia estadística obtenido no sería tan fiable como el obtenido a partir de permutaciones realizadas a partir de cada muestra incluida en el experimento. Para la obtención de conexiones con sentido biológico, pensamos que es esencial una buena elección de la firma molecular introducida para su análisis en el módulo analítico, idealmente introduciendo sólo los genes estadísticamente significativos, resultantes de un análisis de expresión diferencial realizado de manera independiente por el usuario.

También es conveniente señalar que NutriGenomeDB alberga resultados obtenidos a partir de experimentos que podrían no parecer evidentes en un contexto nutricional. Es el caso del estudio de GEO con identificador GSE74212, el cual investiga el efecto del producto natural eusinstielamida B, presente en algunas especies de ascidias. El estudio demuestra que es una molécula tóxica para la proteína topoisomerasa II, encargada de desenredar la molécula de ADN, provoca daños a la molécula de ADN y detiene el crecimiento de células provenientes de cáncer de próstata y mama. Sin embargo, los autores del trabajo no informan sobre las concentraciones de eusinstielamida B presentes naturalmente en esta clase de animales. Por esta razón, utilizado a concentraciones adecuadas, este producto natural podría presentar propiedades interesantes y potencial usabilidad para alimentos funcionales con efectos protectores. De hecho, las ascidias forman parte habitual de la dieta humana en lugares como Japón, Corea, Chile e incluso Europa. Es importante también reconocer que los datos de expresión génica presentes en la base de datos de NutriGenomeDB provienen principalmente de experimentos *in vitro*. No se conoce aún si todos estos compuestos

producen el mismo efecto *in vivo*, tanto a nivel génico como a nivel biológico, y para ello es necesario realizar investigaciones mediante ensayos clínicos específicos.

El uso de herramientas computacionales ha demostrado tener un alto potencial para el descubrimiento de nuevos fármacos y el reposicionamiento de drogas (121-123). Sin embargo, a día de hoy los estudios que emplean metodologías computacionales en el campo nutricional son escasos. El potencial que presentan estos métodos para extraer información sobre el binomio dieta-salud está atrayendo un creciente interés de la comunidad científica, con la perspectiva de hacer de la nutrición personalizada una realidad (124). Por ejemplo, un estudio reciente ha permitido encontrar relaciones entre la dieta y distintas enfermedades gracias a la integración de datos públicos de expresión génica correspondientes a intervenciones nutricionales, enfermedades y drogas (59).

El objetivo de NutriGenomeDB es seguir recopilando nuevos experimentos de nutrigenómica, según vayan apareciendo en las bases de datos públicas de transcriptómica, con el fin de aumentar su capacidad analítica. De esta manera, se dispondrá de un mayor número de firmas moleculares características de alimentos y sus compuestos bioactivos, y será posible identificar potenciales conexiones fiables entre alimentos, sus compuestos bioactivos y determinados medicamentos o drogas, descubrir mecanismos moleculares responsables de tal conexión de manera fiable, e interrogar la base de datos para agrupar los alimentos y sus compuestos bioactivos en función de potenciales propiedades beneficiosas en común.

CONCLUSIONES

Conclusión general

El análisis integrativo de las respuestas genómicas desencadenadas por productos bioactivos alimentarios con propiedades saludables conocidas ofrece interesantes posibilidades para obtener información sobre los efectos moleculares de los nutrientes a nivel celular. A partir de ahí, ha sido posible configurar una herramienta para descubrir nuevas firmas de expresión génica e identificar los mecanismos de acción de los compuestos bioactivos de los alimentos que sustentan sus efectos en la salud. Esta herramienta permite predecir o buscar mecanismos de acción de los componentes de los alimentos y promover así su futura utilización en la formulación de alimentos funcionales o en estrategias de nutrición de precisión.

Por tanto, se confirma la validez de la hipótesis planteada ya que se ha demostrado que la evaluación de datos experimentales de la expresión diferencial de genes permite explicar las propiedades saludables de determinados alimentos, o productos alimentarios, a través de la capacidad de algunos de sus componentes para regular determinadas rutas metabólicas, al afectar a la expresión de los genes implicados.

Conclusiones relativas a los objetivos específicos:

Objetivo específico 1: Análisis experimental del efecto del hidroxitirosol en la expresión de micro ARN's en el hígado de ratón.

1- Una dieta suplementada en HT regula la expresión de determinados miARN's *in vivo* en el hígado de ratón.
2- Los efectos biológicos de la regulación de miARN's por la dieta pueden explicarse a partir del análisis funcional de sus correspondientes genes diana.

Objetivo específico 2: Creación y análisis integrativo de una base de datos a partir de experimentos de nutrigenómica en células humanas.

3- El análisis integrativo del perfil de expresión diferencial de experimentos de nutrigenómica permite identificar alimentos y compuestos bioactivos que comparten propiedades biológicas, independientemente del tipo celular y salvando el efecto "lote".
4- El análisis de agrupamiento jerárquico permite identificar y/o objetivar propiedades beneficiosas de alimentos y sus compuestos bioactivos.

Objetivo específico 3: Desarrollo de una aplicación web para minería de datos en nutrigenómica

5- La aplicación web desarrollada permite conectar experimentos de nutrigenómica y firmas moleculares externas a partir de mecanismos moleculares en común.
6- El Amlodipino, un medicamento utilizado para la hipertensión, y un extracto de Romero comparten el mismo mecanismo molecular de represión de genes implicados en el transporte transmembrana de iones Calcio.
7- La aplicación web NutriGenomeDB posibilita asignar propiedades saludables y beneficiosas a alimentos y sus compuestos bioactivos en base a los perfiles de expresión génica.

BIBLIOGRAFÍA

1. Gibson, T.M., Ferrucci, L.M., Tangrea, J.A., Schatzkin, A. (2010) Epidemiological and clinical studies of nutrition. *Seminars in oncology*, **37**, 282-296.
2. Mahmood, S.S., Levy, D., Vasan, R.S., Wang, T.J. (2014) The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*, **383**, 999-1008.
3. Shekelle, R.B., Shryock, A.M., Paul, O., Lepper, M., Stamler, J., Liu, S., Raynor, W.J., Jr. (1981) Diet, serum cholesterol, and death from coronary heart disease. The Western Electric study. *The New England journal of medicine*, **304**, 65-70.
4. Kushi, L.H., Lew, R.A., Stare, F.J., Ellison, C.R., el Lozy, M., Bourke, G., Daly, L., Graham, I., Hickey, N., Mulcahy, R., et al. (1985) Diet and 20-year mortality from coronary heart disease. The Ireland-Boston Diet-Heart Study. *The New England journal of medicine*, **312**, 811-818.
5. Sofi, F., Abbate, R., Gensini, G.F., Casini, A. (2010) Accruing evidence on benefits of adherence to the Mediterranean diet on health: an updated systematic review and meta-analysis. *The American journal of clinical nutrition*, **92**, 1189-1196.
6. Estruch, R., Ros, E., Salas-Salvado, J., Covas, M.I., Corella, D., Aros, F., Gomez-Gracia, E., Ruiz-Gutierrez, V., Fiol, M., Lapetra, J., Lamuela-Raventos, R.M., Serra-Majem, L., Pinto, X., Basora, J., Munoz, M.A., Sorli, J.V., Martinez, J.A., Fito, M., Gea, A., Hernan, M.A., Martinez-Gonzalez, M.A. (2018) Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. *The New England journal of medicine*, **378**, e34.
7. Vincent-Baudry, S., Defoort, C., Gerber, M., Bernard, M.C., Verger, P., Helal, O., Portugal, H., Planells, R., Grolier, P., Amiot-Carlin, M.J., Vague, P., Lairon, D. (2005) The Medi-RIVAGE study: reduction of cardiovascular disease risk factors after a 3-mo intervention with a Mediterranean-type diet or a low-fat diet. *The American journal of clinical nutrition*, **82**, 964-971.
8. Esposito, K., Marfella, R., Ciotola, M., Di Palo, C., Giugliano, F., Giugliano, G., D'Armiento, M., D'Andrea, F., Giugliano, D. (2004) Effect of a mediterranean-style diet on endothelial dysfunction and markers of vascular inflammation in the metabolic syndrome: a randomized trial. *Jama*, **292**, 1440-1446.
9. Appel, L.J., Moore, T.J., Obarzanek, E., Vollmer, W.M., Svetkey, L.P., Sacks, F.M., Bray, G.A., Vogt, T.M., Cutler, J.A., Windhauser, M.M., Lin, P.H., Karanja, N. (1997) A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. *The New England journal of medicine*, **336**, 1117-1124.
10. McKeown, N.M., Meigs, J.B., Liu, S., Saltzman, E., Wilson, P.W., Jacques, P.F. (2004) Carbohydrate nutrition, insulin resistance, and the prevalence of the metabolic syndrome in the Framingham Offspring Cohort. *Diabetes care*, **27**, 538-546.
11. Di Daniele, N., Noce, A., Vidiri, M.F., Moriconi, E., Marrone, G., Annicchiarico-Petruzzelli, M., D'Urso, G., Tesauro, M., Rovella, V., De Lorenzo, A. (2017) Impact of Mediterranean diet on metabolic syndrome, cancer and longevity. *Oncotarget*, **8**, 8947-8979.
12. Couto, E., Boffetta, P., Lagiou, P., Ferrari, P., Buckland, G., Overvad, K., Dahm, C.C., Tjonneland, A., Olsen, A., Clavel-Chapelon, F., Boutron-Ruault, M.C., Cottet, V., Trichopoulos, D., Naska, A., Benetou, V., Kaaks, R., Rohrmann, S., Boeing, H., von Ruesten, A., Panico, S., Pala, V., Vineis, P., Palli, D., Tumino, R., May, A., Peeters, P.H., Bueno-de-Mesquita, H.B., Buchner, F.L., Lund, E., Skeie, G., Engeset, D., Gonzalez, C.A., Navarro, C., Rodriguez, L., Sanchez, M.J., Amiano, P., Barricarte, A., Hallmans, G., Johansson, I., Manjer, J., Wirfart, E., Allen, N.E., Crowe, F., Khaw, K.T., Wareham, N., Moskal, A., Slimani, N., Jenab, M., Romaguera, D., Mouw, T., Norat, T., Riboli, E., Trichopoulou, A. (2011) Mediterranean dietary pattern and cancer risk in the EPIC cohort. *British journal of cancer*, **104**, 1493-1499.
13. Meyerhardt, J.A., Niedzwiecki, D., Hollis, D., Saltz, L.B., Hu, F.B., Mayer, R.J., Nelson, H., Whittom, R., Hantel, A., Thomas, J., Fuchs, C.S. (2007) Association of dietary patterns with cancer recurrence and survival in patients with stage III colon cancer. *Jama*, **298**, 754-764.
14. Schmidt, M., Pfetzer, N., Schwab, M., Strauss, I., Kammerer, U. (2011) Effects of a ketogenic diet on the quality of life in 16 patients with advanced cancer: A pilot trial. *Nutrition & metabolism*, **8**, 54.
15. Beresford, S.A., Johnson, K.C., Ritenbaugh, C., Lasser, N.L., Snetselaar, L.G., Black, H.R., Anderson, G.L., Assaf, A.R., Bassford, T., Bowen, D., Brunner, R.L., Brzyski, R.G., Caan, B., Chlebowski, R.T., Gass, M., Harrigan, R.C., Hays, J., Heber, D., Heiss, G., Hendrix, S.L., Howard, B.V., Hsia, J., Hubbell, F.A., Jackson, R.D., Kotchen, J.M., Kuller, L.H., LaCroix, A.Z., Lane, D.S., Langer, R.D., Lewis, C.E., Manson, J.E., Margolis, K.L., Mossavar-Rahmani, Y., Ockene, J.K., Parker, L.M., Perri, M.G., Phillips, L., Prentice, R.L., Robbins, J., Rossouw, J.E., Sarto, G.E., Stefanick, M.L., Van Horn, L., Vitolins, M.Z., Wactawski-Wende, J., Wallace, R.B., Whitlock, E. (2006) Low-fat dietary pattern and risk of

colorectal cancer: the Women's Health Initiative Randomized Controlled Dietary Modification Trial. *Jama*, **295**, 643-654.
16. Willett, W. (2008) Nutrition and cancer: the search continues. *Nutrition and cancer*, **60**, 557-559.
17. Harris, W.S., Bulchandani, D. (2006) Why do omega-3 fatty acids lower serum triglycerides? *Current opinion in lipidology*, **17**, 387-393.
18. James, M., Proudman, S., Cleland, L. (2010) Fish oil and rheumatoid arthritis: past, present and future. *The Proceedings of the Nutrition Society*, **69**, 316-323.
19. Kliewer, S.A., Sundseth, S.S., Jones, S.A., Brown, P.J., Wisely, G.B., Koble, C.S., Devchand, P., Wahli, W., Willson, T.M., Lenhard, J.M., Lehmann, J.M. (1997) Fatty acids and eicosanoids regulate gene expression through direct interactions with peroxisome proliferator-activated receptors alpha and gamma. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 4318-4323.
20. Mishra, A., Chaudhary, A., Sethi, S. (2004) Oxidized omega-3 fatty acids inhibit NF-kappaB activation via a PPARalpha-dependent pathway. *Arteriosclerosis, thrombosis, and vascular biology*, **24**, 1621-1627.
21. Mozaffarian, D., Wu, J.H. (2011) Omega-3 fatty acids and cardiovascular disease: effects on risk factors, molecular pathways, and clinical events. *Journal of the American College of Cardiology*, **58**, 2047-2067.
22. Gonzalez-Vallinas, M., Molina, S., Vicente, G., Zarza, V., Martin-Hernandez, R., Garcia-Risco, M.R., Fornari, T., Reglero, G., Ramirez de Molina, A. (2014) Expression of microRNA-15b and the glycosyltransferase GCNT3 correlates with antitumor efficacy of Rosemary diterpenes in colon and pancreatic cancer. *PloS one*, **9**, e98556.
23. Park, O.J., Surh, Y.J. (2004) Chemopreventive potential of epigallocatechin gallate and genistein: evidence from epidemiological and laboratory studies. *Toxicology letters*, **150**, 43-56.
24. Shankar, S., Chen, Q., Srivastava, R.K. (2008) Inhibition of PI3K/AKT and MEK/ERK pathways act synergistically to enhance antiangiogenic effects of EGCG through activation of FOXO transcription factor. *Journal of molecular signaling*, **3**, 7.
25. Singh, B.N., Shankar, S., Srivastava, R.K. (2011) Green tea catechin, epigallocatechin-3-gallate (EGCG): mechanisms, perspectives and clinical applications. *Biochemical pharmacology*, **82**, 1807-1821.
26. Corella, D., Coltell, O., Macian, F., Ordovas, J.M. (2018) Advances in Understanding the Molecular Basis of the Mediterranean Diet Effect. *Annual review of food science and technology*, **9**, 227-249.
27. Comalada, M., Ballester, I., Bailon, E., Sierra, S., Xaus, J., Galvez, J., de Medina, F.S., Zarzuelo, A. (2006) Inhibition of pro-inflammatory markers in primary bone marrow-derived mouse macrophages by naturally occurring flavonoids: analysis of the structure-activity relationship. *Biochemical pharmacology*, **72**, 1010-1021.
28. Schon, E.A., Przedborski, S. (2011) Mitochondria: the next (neurode)generation. *Neuron*, **70**, 1033-1053.
29. Galleano, M., Verstraeten, S.V., Oteiza, P.I., Fraga, C.G. (2010) Antioxidant actions of flavonoids: thermodynamic and kinetic analysis. *Archives of biochemistry and biophysics*, **501**, 23-30.
30. Egert, S., Bosy-Westphal, A., Seiberl, J., Kurbitz, C., Settler, U., Plachta-Danielzik, S., Wagner, A.E., Frank, J., Schrezenmeir, J., Rimbach, G., Wolffram, S., Muller, M.J. (2009) Quercetin reduces systolic blood pressure and plasma oxidised low-density lipoprotein concentrations in overweight subjects with a high-cardiovascular disease risk phenotype: a double-blinded, placebo-controlled cross-over study. *The British journal of nutrition*, **102**, 1065-1074.
31. Bharrhan, S., Chopra, K., Arora, S.K., Toor, J.S., Rishi, P. (2012) Down-regulation of NF-kappaB signalling by polyphenolic compounds prevents endotoxin-induced liver injury in a rat model. *Innate immunity*, **18**, 70-79.
32. Weng, C.J., Chen, M.J., Yeh, C.T., Yen, G.C. (2011) Hepatoprotection of quercetin against oxidative stress by induction of metallothionein expression through activating MAPK and PI3K pathways and enhancing Nrf2 DNA-binding activity. *New biotechnology*, **28**, 767-777.
33. Zhao, S.G., Li, Q., Liu, Z.X., Wang, J.J., Wang, X.X., Qin, M., Wen, Q.S. (2011) Curcumin attenuates insulin resistance in hepatocytes by inducing Nrf2 nuclear translocation. *Hepato-gastroenterology*, **58**, 2106-2111.
34. Mueller, L., Boehm, V. (2011) Antioxidant activity of beta-carotene compounds in different in vitro assays. *Molecules*, **16**, 1055-1069.

35. Howitz, K.T., Bitterman, K.J., Cohen, H.Y., Lamming, D.W., Lavu, S., Wood, J.G., Zipkin, R.E., Chung, P., Kisielewski, A., Zhang, L.L., Scherer, B., Sinclair, D.A. (2003) Small molecule activators of sirtuins extend Saccharomyces cerevisiae lifespan. *Nature*, **425**, 191-196.
36. Fullerton, M.D., Steinberg, G.R. (2010) SIRT1 takes a backseat to AMPK in the regulation of insulin sensitivity by resveratrol. *Diabetes*, **59**, 551-553.
37. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., Enright, A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, **34**, D140-144.
38. Vasudevan, S., Tong, Y., Steitz, J.A. (2007) Switching from repression to activation: microRNAs can up-regulate translation. *Science*, **318**, 1931-1934.
39. Place, R.F., Li, L.C., Pookot, D., Noonan, E.J., Dahiya, R. (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 1608-1613.
40. Rayner, K.J., Suarez, Y., Davalos, A., Parathath, S., Fitzgerald, M.L., Tamehiro, N., Fisher, E.A., Moore, K.J., Fernandez-Hernando, C. (2010) MiR-33 contributes to the regulation of cholesterol homeostasis. *Science*, **328**, 1570-1573.
41. Gil-Zamorano, J., Martin, R., Daimiel, L., Richardson, K., Giordano, E., Nicod, N., Garcia-Carrasco, B., Soares, S.M., Iglesias-Gutierrez, E., Lasuncion, M.A., Sala-Vila, A., Ros, E., Ordovas, J.M., Visioli, F., Davalos, A. (2014) Docosahexaenoic acid modulates the enterocyte Caco-2 cell expression of microRNAs involved in lipid metabolism. *The Journal of nutrition*, **144**, 575-585.
42. Aridi, Y.S., Walker, J.L., Wright, O.R.L. (2017) The Association between the Mediterranean Dietary Pattern and Cognitive Health: A Systematic Review. *Nutrients*, **9**.
43. Crespo, M.C., Tome-Carneiro, J., Gomez-Coronado, D., Burgos-Ramos, E., Garcia-Serrano, A., Martin-Hernandez, R., Baliyan, S., Fontecha, J., Venero, C., Davalos, A., Visioli, F. (2018) Modulation of miRNA expression in aged rat hippocampus by buttermilk and krill oil. *Scientific reports*, **8**, 3993.
44. Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., Li, J., Bian, Z., Liang, X., Cai, X., Yin, Y., Wang, C., Zhang, T., Zhu, D., Zhang, D., Xu, J., Chen, Q., Ba, Y., Liu, J., Wang, Q., Chen, J., Wang, J., Wang, M., Zhang, Q., Zhang, J., Zen, K., Zhang, C.Y. (2012) Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell research*, **22**, 107-126.
45. Mu, J., Zhuang, X., Wang, Q., Jiang, H., Deng, Z.B., Wang, B., Zhang, L., Kakar, S., Jun, Y., Miller, D., Zhang, H.G. (2014) Interspecies communication between plant and mouse gut host cells through edible plant derived exosome-like nanoparticles. *Molecular nutrition & food research*, **58**, 1561-1573.
46. Arpon, A., Riezu-Boj, J.I., Milagro, F.I., Marti, A., Razquin, C., Martinez-Gonzalez, M.A., Corella, D., Estruch, R., Casas, R., Fito, M., Ros, E., Salas-Salvado, J., Martinez, J.A. (2016) Adherence to Mediterranean diet is associated with methylation changes in inflammation-related genes in peripheral blood cells. *Journal of physiology and biochemistry*, **73**, 445-455.
47. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**, R80.
48. Bumgarner, R. (2013) Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology*, **Chapter 22**, Unit 22 21.
49. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.
50. Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, **3**, Article3.
51. Tusher, V.G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.
52. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., Golani, I. (2001) Controlling the false discovery rate in behavior genetics research. *Behavioural brain research*, **125**, 279-284.
53. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Roder,

M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., Gingeras, T.R. (2012) Landscape of transcription in human cells. *Nature*, **489**, 101-108.

54. Hatem, A., Bozdag, D., Toland, A.E., Catalyurek, U.V. (2013) Benchmarking short sequence mapping tools. *BMC bioinformatics*, **14**, 184.
55. Anders, S., Pyl, P.T., Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166-169.
56. Robinson, M.D., McCarthy, D.J., Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.
57. Robinson, M.D., Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, **11**, R25.
58. Clough, E., Barrett, T. (2016) The Gene Expression Omnibus Database. *Methods Mol Biol*, **1418**, 93-110.
59. Zheng, T., Ni, Y., Li, J., Chow, B.K.C., Panagiotou, G. (2017) Designing Dietary Recommendations Using System Level Interactomics Analysis and Network-Based Inference. *Frontiers in physiology*, **8**, 753.
60. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929-1935.
61. Wang, G., Ye, Y., Yang, X., Liao, H., Zhao, C., Liang, S. (2011) Expression-based in silico screening of candidate therapeutic compounds for lung adenocarcinoma. *PloS one*, **6**, e14573.
62. Wei, G., Twomey, D., Lamb, J., Schlis, K., Agarwal, J., Stam, R.W., Opferman, J.T., Sallan, S.E., den Boer, M.L., Pieters, R., Golub, T.R., Armstrong, S.A. (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer cell*, **10**, 331-342.
63. Kibble, M., Khan, S.A., Saarinen, N., Iorio, F., Saez-Rodriguez, J., Makela, S., Aittokallio, T. (2016) Transcriptional response networks for elucidating mechanisms of action of multitargeted agents. *Drug discovery today*, **21**, 1063-1075.
64. Williams, G. (2013) SPIEDw: a searchable platform-independent expression database web tool. *BMC genomics*, **14**, 765.
65. Zinman, G.E., Naiman, S., Kanfi, Y., Cohen, H., Bar-Joseph, Z. (2013) ExpressionBlast: mining large, unstructured expression databases. *Nature methods*, **10**, 925-926.
66. Wang, Z., Monteiro, C.D., Jagodnik, K.M., Fernandez, N.F., Gundersen, G.W., Rouillard, A.D., Jenkins, S.L., Feldmann, A.S., Hu, K.S., McDermott, M.G., Duan, Q., Clark, N.R., Jones, M.R., Kou, Y., Goff, T., Woodland, H., Amaral, F.M.R., Szeto, G.L., Fuchs, O., Schussler-Fiorenza Rose, S.M., Sharma, S., Schwartz, U., Bausela, X.B., Szymkiewicz, M., Maroulis, V., Salykin, A., Barra, C.M., Kruth, C.D., Bongio, N.J., Mathur, V., Todoric, R.D., Rubin, U.E., Malatras, A., Fulp, C.T., Galindo, J.A., Motiejunaite, R., Juschke, C., Dishuck, P.C., Lahl, K., Jafari, M., Aibar, S., Zaravinos, A., Steenhuizen, L.H., Allison, L.R., Gamallo, P., de Andres Segura, F., Dae Devlin, T., Perez-Garcia, V., Ma'ayan, A. (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature communications*, **7**, 12846.
67. Setoain, J., Franch, M., Martinez, M., Tabas-Madrid, D., Sorzano, C.O., Bakker, A., Gonzalez-Couto, E., Elvira, J., Pascual-Montano, A. (2015) NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic acids research*, **43**, W193-199.
68. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

69. Zhang, S.D., Gant, T.W. (2008) A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC bioinformatics*, **9**, 258.
70. Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R., Golub, T.R. (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834-838.
71. Kluger, Y., Basri, R., Chang, J.T., Gerstein, M. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, **13**, 703-716.
72. Langmead, B., Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357-359.
73. Sticht, C., De La Torre, C., Parveen, A., Gretz, N. (2018) miRWalk: An online resource for prediction of microRNA binding sites. *PloS one*, **13**, e0206239.
74. Mi, H., Muruganujan, A., Thomas, P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, **41**, D377-386.
75. Tabas-Madrid, D., Nogales-Cadenas, R., Pascual-Montano, A. (2012) GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic acids research*, **40**, W478-483.
76. Fernandez, N.F., Gundersen, G.W., Rahman, A., Grimes, M.L., Rikova, K., Hornbeck, P., Ma'ayan, A. (2017) Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data*, **4**, 170151.
77. Ahsan, H., Ahad, A., Iqbal, J., Siddiqui, W.A. (2014) Pharmacological potential of tocotrienols: a review. *Nutrition & metabolism*, **11**, 52.
78. Contos, J.J., Ishii, I., Chun, J. (2000) Lysophosphatidic acid receptors. *Molecular pharmacology*, **58**, 1188-1196.
79. Fuhr, L., Rousseau, M., Plauth, A., Schroeder, F.C., Sauer, S. (2015) Amorfrutins Are Natural PPARgamma Agonists with Potent Anti-inflammatory Properties. *Journal of natural products*, **78**, 1160-1164.
80. Richard, D., Kefi, K., Barbe, U., Bausero, P., Visioli, F. (2008) Polyunsaturated fatty acids as antioxidants. *Pharmacological research*, **57**, 451-455.
81. Quirit, J.G., Lavrenov, S.N., Poindexter, K., Xu, J., Kyauk, C., Durkin, K.A., Aronchik, I., Tomasiak, T., Solomatin, Y.A., Preobrazhenskaya, M.N., Firestone, G.L. (2017) Indole-3-carbinol (I3C) analogues are potent small molecule inhibitors of NEDD4-1 ubiquitin ligase activity that disrupt proliferation of human melanoma cells. *Biochemical pharmacology*, **127**, 13-27.
82. Szarc vel Szic, K., Op de Beeck, K., Ratman, D., Wouters, A., Beck, I.M., Declerck, K., Heyninck, K., Fransen, E., Bracke, M., De Bosscher, K., Lardon, F., Van Camp, G., Vanden Berghe, W. (2014) Pharmacological levels of Withaferin A (Withania somnifera) trigger clinically relevant anticancer effects specific to triple negative breast cancer cells. *PloS one*, **9**, e87850.
83. Agyeman, A.S., Chaerkady, R., Shaw, P.G., Davidson, N.E., Visvanathan, K., Pandey, A., Kensler, T.W. (2012) Transcriptomic and proteomic profiling of KEAP1 disrupted and sulforaphane-treated human breast epithelial cells reveals common expression profiles. *Breast cancer research and treatment*, **132**, 175-187.
84. Varoni, E.M., Lo Faro, A.F., Sharifi-Rad, J., Iriti, M. (2016) Anticancer Molecular Mechanisms of Resveratrol. *Frontiers in nutrition*, **3**, 8.
85. Forsyth, C.B., Farhadi, A., Jakate, S.M., Tang, Y., Shaikh, M., Keshavarzian, A. (2009) Lactobacillus GG treatment ameliorates alcohol-induced intestinal oxidative stress, gut leakiness, and liver injury in a rat model of alcoholic steatohepatitis. *Alcohol*, **43**, 163-172.
86. Zheng, Z., Yu, S., Zhang, W., Peng, Y., Pu, M., Kang, T., Zeng, J., Yu, Y., Li, G. (2017) Genistein attenuates monocrotaline-induced pulmonary arterial hypertension in rats by activating PI3K/Akt/eNOS signaling. *Histology and histopathology*, **32**, 35-41.
87. Sun, L., Zhao, T., Ju, T., Wang, X., Li, X., Wang, L., Zhang, L., Yu, G. (2015) A Combination of Intravenous Genistein Plus Mg2+ Enhances Antihypertensive Effects in SHR by Endothelial Protection and BKCa Channel Inhibition. *American journal of hypertension*, **28**, 1114-1120.
88. Matori, H., Umar, S., Nadadur, R.D., Sharma, S., Partow-Navid, R., Afkhami, M., Amjedi, M., Eghbali, M. (2012) Genistein, a soy phytoestrogen, reverses severe pulmonary hypertension and prevents right heart failure in rats. *Hypertension*, **60**, 425-430.
89. Ordovas, J.M., Corella, D. (2004) Genes, diet and plasma lipids: the evidence from observational studies. *World review of nutrition and dietetics*, **93**, 41-76.

90. Simopoulos, A.P. (2010) Genetic variants in the metabolism of omega-6 and omega-3 fatty acids: their role in the determination of nutritional requirements and chronic disease risk. *Exp Biol Med (Maywood)*, **235**, 785-795.
91. Alkhatib, A., Tsang, C., Tuomilehto, J. (2018) Olive Oil Nutraceuticals in the Prevention and Management of Diabetes: From Molecules to Lifestyle. *International journal of molecular sciences*, **19**.
92. Logan, J., Bourassa, M.W. (2018) The rationale for a role for diet and nutrition in the prevention and treatment of cancer. *Eur J Cancer Prev*, **27**, 406-410.
93. Pan, M.H., Wu, J.C., Ho, C.T., Lai, C.S. (2018) Antiobesity molecular mechanisms of action: Resveratrol and pterostilbene. *Biofactors*, **44**, 50-60.
94. Milenkovic, D., Deval, C., Gouranton, E., Landrier, J.F., Scalbert, A., Morand, C., Mazur, A. (2012) Modulation of miRNA expression by dietary polyphenols in apoE deficient mice: a new mechanism of the action of polyphenols. *PloS one*, **7**, e29837.
95. Baselga-Escudero, L., Blade, C., Ribas-Latre, A., Casanova, E., Salvado, M.J., Arola, L., Arola-Arnal, A. (2014) Chronic supplementation of proanthocyanidins reduces postprandial lipemia and liver miR-33a and miR-122 levels in a dose-dependent manner in healthy rats. *The Journal of nutritional biochemistry*, **25**, 151-156.
96. Nunez-Sanchez, M.A., Davalos, A., Gonzalez-Sarrias, A., Casas-Agustench, P., Visioli, F., Monedero-Saiz, T., Garcia-Talavera, N.V., Gomez-Sanchez, M.B., Sanchez-Alvarez, C., Garcia-Albert, A.M., Rodriguez-Gil, F.J., Ruiz-Marin, M., Pastor-Quirante, F.A., Martinez-Diaz, F., Tomas-Barberan, F.A., Garcia-Conesa, M.T., Espin, J.C. (2015) MicroRNAs expression in normal and malignant colon tissues as biomarkers of colorectal cancer and in response to pomegranate extracts consumption: Critical issues to discern between modulatory effects and potential artefacts. *Molecular nutrition & food research*, **59**, 1973-1986.
97. D'Adamo, S., Cetrullo, S., Guidotti, S., Borzi, R.M., Flamigni, F. (2017) Hydroxytyrosol modulates the levels of microRNA-9 and its target sirtuin-1 thereby counteracting oxidative stress-induced chondrocyte death. *Osteoarthritis and cartilage*, **25**, 600-610.
98. Bigagli, E., Cinci, L., Paccosi, S., Parenti, A., D'Ambrosio, M., Luceri, C. (2017) Nutritionally relevant concentrations of resveratrol and hydroxytyrosol mitigate oxidative burst of human granulocytes and monocytes and the production of pro-inflammatory mediators in LPS-stimulated RAW 264.7 macrophages. *International immunopharmacology*, **43**, 147-155.
99. Tome-Carneiro, J., Crespo, M.C., Iglesias-Gutierrez, E., Martin, R., Gil-Zamorano, J., Tomas-Zapico, C., Burgos-Ramos, E., Correa, C., Gomez-Coronado, D., Lasuncion, M.A., Herrera, E., Visioli, F., Davalos, A. (2016) Hydroxytyrosol supplementation modulates the expression of miRNAs in rodents and in humans. *The Journal of nutritional biochemistry*, **34**, 146-155.
100. Kornfeld, J.W., Baitzel, C., Konner, A.C., Nicholls, H.T., Vogt, M.C., Herrmanns, K., Scheja, L., Haumaitre, C., Wolf, A.M., Knippschild, U., Seibler, J., Cereghini, S., Heeren, J., Stoffel, M., Bruning, J.C. (2013) Obesity-induced overexpression of miR-802 impairs glucose metabolism through silencing of Hnf1b. *Nature*, **494**, 111-115.
101. Sansom, S.E., Nuovo, G.J., Martin, M.M., Kotha, S.R., Parinandi, N.L., Elton, T.S. (2010) miR-802 regulates human angiotensin II type 1 receptor expression in intestinal epithelial C2BBe1 cells. *American journal of physiology. Gastrointestinal and liver physiology*, **299**, G632-642.
102. Li, H.T., Zhang, H., Chen, Y., Liu, X.F., Qian, J. (2015) MiR-423-3p enhances cell growth through inhibition of p21Cip1/Waf1 in colorectal cancer. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology*, **37**, 1044-1054.
103. Lin, J., Huang, S., Wu, S., Ding, J., Zhao, Y., Liang, L., Tian, Q., Zha, R., Zhan, R., He, X. (2011) MicroRNA-423 promotes cell growth and regulates G(1)/S transition by targeting p21Cip1/Waf1 in hepatocellular carcinoma. *Carcinogenesis*, **32**, 1641-1647.
104. Chen, J., Yu, Y., Li, S., Liu, Y., Zhou, S., Cao, S., Yin, J., Li, G. (2017) MicroRNA-30a ameliorates hepatic fibrosis by inhibiting Beclin1-mediated autophagy. *Journal of cellular and molecular medicine*, **21**, 3679-3692.
105. Li, L., Kang, L., Zhao, W., Feng, Y., Liu, W., Wang, T., Mai, H., Huang, J., Chen, S., Liang, Y., Han, J., Xu, X., Ye, Q. (2017) miR-30a-5p suppresses breast tumor growth and metastasis through inhibition of LDHA-mediated Warburg effect. *Cancer letters*, **400**, 89-98.
106. Jiang, W., Liu, J., Dai, Y., Zhou, N., Ji, C., Li, X. (2015) MiR-146b attenuates high-fat diet-induced non-alcoholic steatohepatitis in mice. *Journal of gastroenterology and hepatology*, **30**, 933-943.

107. Li, C., Miao, R., Liu, S., Wan, Y., Zhang, S., Deng, Y., Bi, J., Qu, K., Zhang, J., Liu, C. (2017) Down-regulation of miR-146b-5p by long noncoding RNA MALAT1 in hepatocellular carcinoma promotes cancer growth and metastasis. *Oncotarget*, **8**, 28683-28695.
108. Betel, D., Wilson, M., Gabow, A., Marks, D.S., Sander, C. (2008) The microRNA.org resource: targets and expression. *Nucleic acids research*, **36**, D149-153.
109. Agarwal, V., Bell, G.W., Nam, J.W., Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**.
110. Betel, D., Koppal, A., Agius, P., Sander, C., Leslie, C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology*, **11**, R90.
111. Karagkouni, D., Paraskevopoulou, M.D., Chatzopoulos, S., Vlachos, I.S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S., Skoufos, G., Vergoulis, T., Dalamagas, T., Hatzigeorgiou, A.G. (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic acids research*, **46**, D239-D245.
112. Ni, W.J., Leng, X.M. (2015) Dynamic miRNA-mRNA paradigms: New faces of miRNAs. *Biochemistry and biophysics reports*, **4**, 337-341.
113. Lee, E.J., Baek, M., Gusev, Y., Brackett, D.J., Nuovo, G.J., Schmittgen, T.D. (2008) Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors. *RNA*, **14**, 35-42.
114. Zhang, W., He, W., Shi, Y., Gu, H., Li, M., Liu, Z., Feng, Y., Zheng, N., Xie, C., Zhang, Y. (2016) High Expression of KIF20A Is Associated with Poor Overall Survival and Tumor Progression in Early-Stage Cervical Squamous Cell Carcinoma. *PloS one*, **11**, e0167449.
115. Oh, M., Ahn, J., Lee, T., Jang, G., Park, C., Yoon, Y. (2017) Drug voyager: a computational platform for exploring unintended drug action. *BMC bioinformatics*, **18**, 131.
116. Hassani, F.V., Shirani, K., Hosseinzadeh, H. (2016) Rosemary (Rosmarinus officinalis) as a potential therapeutic plant in metabolic syndrome: a review. *Naunyn-Schmiedeberg's archives of pharmacology*, **389**, 931-949.
117. Apostolidis, E., Kwon, Y.I., Shetty, K. (2006) Potential of cranberry-based herbal synergies for diabetes and hypertension management. *Asia Pacific journal of clinical nutrition*, **15**, 433-441.
118. Neves, J.A., Oliveira, R.C.M. (2018) Pharmacological and biotechnological advances with Rosmarinus officinalis L. *Expert opinion on therapeutic patents*, **28**, 399-413.
119. Sureda, A., Sanches Silva, A., Sanchez-Machado, D.I., Lopez-Cervantes, J., Daglia, M., Nabavi, S.F., Nabavi, S.M. (2017) Hypotensive effects of genistein: From chemistry to medicine. *Chemico-biological interactions*, **268**, 37-46.
120. Tamayo, P., Steinhardt, G., Liberzon, A., Mesirov, J.P. (2016) The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical methods in medical research*, **25**, 472-487.
121. Chaudhari, R., Tan, Z., Huang, B., Zhang, S. (2017) Computational polypharmacology: a new paradigm for drug discovery. *Expert opinion on drug discovery*, **12**, 279-291.
122. Napolitano, F., Carrella, D., Mandriani, B., Pisonero-Vaquero, S., Sirci, F., Medina, D.L., Brunetti-Pierri, N., di Bernardo, D. (2018) gene2drug: a computational tool for pathway-based rational drug repositioning. *Bioinformatics*, **34**, 1498-1505.
123. Wan, F., Hong, L., Xiao, A., Jiang, T., Zeng, J. (2019) NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics*, **35**, 104-111.
124. Badimon, L., Vilahur, G., Padro, T. (2017) Systems biology approaches to understand the effects of nutrition and promote health. *British journal of clinical pharmacology*, **83**, 38-45.

ANEXO

Journal of Functional Foods 42 (2018) 380–386

Contents lists available at ScienceDirect

Journal of Functional Foods

journal homepage: www.elsevier.com/locate/jff

Data mining of nutrigenomics experiments: Identification of a cancer protective gene signature

Roberto Martín-Hernández[a,*], Guillermo Reglero[b,c], Alberto Dávalos[d]

[a] Bioinformatics Unit, IMDEA Food Institute, CEI UAM + CSIC, Madrid 28049, Spain
[b] Sección Departamental de Ciencias de la Alimentación, Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid 28049, Spain
[c] Laboratory of Food Products for Precision Nutrition, IMDEA Food Institute, CEI UAM + CSIC, Madrid 28049, Spain
[d] Laboratory of Epigenetics of Lipid Metabolism, IMDEA Food Institute, CEI UAM + CSIC, Madrid 28049, Spain

ARTICLE INFO

Keywords:
Nutrigenomics
Microarrays
Clustering
Gene signature
Anticancer

ABSTRACT

Regular consumption of certain foods has shown beneficial effects on cardiometabolic health. However, it is not clear by which molecular mechanisms they may exert their beneficial effects. Many genomic experiments available in public databases have generated gene expression data following the treatment of human cells with different food nutrients. Exploration of such data offers great possibilities for gaining insights into the molecular effects of nutrients at cellular level. In this work, we explored the genomic responses triggered by food bioactive compounds with well-known healthy properties. We show that human cell lines treated with different food compounds tend to cluster in a cell type dependent manner based on gene expression, with an influence of the physiological attributes of cells. Finally, we identify a genomic signature of 18 genes implicated in cell cycle, which may characterize a protective effect of certain food compounds against cancer. Our data provides evidence that nutrigenomic studies found in public databases can be used to discover novel signatures of gene expression and identify common mechanism of actions of food bioactive compounds.

1. Introduction

Nutritional genomics, also known as nutrigenomics, is a relatively new science which explores the effects of nutrients on the genome, proteome and metabolome. Whereas the idea of modulating human health by food intake is a millennial concept, there are great expectations on the tremendous potential this science may have to change the future of dietary guidelines in order to improve health and hence to build up a precision nutrition era (DeBusk, Fogarty, Ordovas, & Kornman, 2005).

A functional food has been defined as “any modified food or food ingredient that may provide a health benefit beyond that of the traditional nutrients it contains” (Snetselaar, 1994). During the last 20 years there have been substantial efforts to identify bioactive compounds in food which might be associated with beneficial biological activities. For example compounds such as long-chain polyunsaturated fatty acids (n-3 PUFAS), which consumption has been associated with a reduced risk of cardiovascular disease, are known to act as ligands for cellular receptors to trigger a signaling cascade that inhibits the expression of proinflammatory genes (Ferguson, 2009). Also, many natural products, extracted from foods used in human diet, have shown great potential as anti-proliferative agents on cultured cancer human cells (Gonzalez-Vallinas et al., 2013; Ramirez de Molina et al., 2015). Indeed, a wide range of drugs for treating diseases such as diabetes and cancer are derived from natural products. Interestingly, some food compounds have proved their ability to interact with the epigenome, thus modifying microRNA expression (Gil-Zamorano et al., 2014). However, the molecular mechanisms by which food bioactive compounds exert their beneficial effects are still not well understood.

Omics technologies are widely adopted to study the expression of thousands of genes and proteins at a time. These technologies generate a vast amount of gene expression data that accumulates in public repositories such as the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013). Whereas these data remains unclassified by phenotype or experimental condition, the user interface allows easily querying and mining the database for experiments.

Other databases such as the Broad Institute's Connectivity Map (CMap) (Lamb et al., 2006) collect highly specific expression data from cell lines treated with drugs and other chemicals. Such type of transcriptomic data has previously been utilized to establish functional

Abbreviations: GEO, NCBI Gene Expression Omnibus; GES, Gene Expression Signature; TCT, Tocotrienols; LB, Lactobacillus; AMF, Amorfrutin; I3C, Indole-3-carbinol; RSM, Rosemary; CA, Carnosic Acid; WFNA, Withaferin A; SFN, Sulforaphane; RVT, Resveratrol; TRVT, Transresveratrol; FC, Fold change

* Corresponding author at: Bioinformatics Unit, IMDEA Food Institute, CEI UAM + CSIC, Ctra. De Canto Blanco 8, E-28049 Madrid, Spain.
E-mail address: roberto.martin@imdea.org (R. Martín-Hernández).

https://doi.org/10.1016/j.jff.2018.01.021
Received 30 October 2017; Received in revised form 19 January 2018; Accepted 20 January 2018
Available online 03 February 2018
1756-4646/

connections between drugs, genes and diseases using computational approaches. From a transcriptomic point of view, mathematical models have already been applied on gene expressions data for the identification of pathway responsiveness to drugs (Pratanwanich & Lio, 2014). Another approach allows the generation of a list of drugs triggering a similar gene expression pattern at cellular level (Lee et al., 2012) and thus possibly sharing a common mechanism of action. From a disease point of view, other approaches consider that a particular gene expression signature (GES) related to a disease might be reverted using a drug which triggers an opposite GES (Setoain et al., 2015), showing promising opportunities within the drug repositioning field (Jia et al., 2016). Artificial intelligence, and specifically deep learning algorithms, has been applied on large transcriptional response data sets with the aim of classifying various drugs to therapeutic categories solely based on their transcriptional profiles (Aliper et al., 2016). However, to the best of our knowledge there is no evidence about such approaches applied to the emerging field of nutrigenomic studies, seeking to investigate the effect of food and nutrients on gene expression.

We extracted from GEO repository all the available experiments related to nutrigenomics in human cells to survey the gene expression patterns. The correlation of gene expression patterns can show potential connections between bioactive compounds, indicating that they may share a common mechanism of action, and allowing the discovery of new potential therapeutic molecules (Lamb et al., 2006). Here we present a comprehensive data mining analysis of a set of nutrigenomics experiments extracted from GEO database. The assessment of human cell's gene expression cultured in vitro after treatment with bioactive compounds obtained from food should lead to a better characterization of the molecular mechanisms that confer a beneficial effect to certain food products.

2. Materials and methods

2.1. Data collection and analysis

Studies corresponding to nutrigenomics were identified from GEO database. Specific queries were launched containing words such as “nutrient”, “nutrition”, “natural product”, “extract” and “phytochemical”. For data corresponding to Affymetrix platforms, raw data was downloaded and normalized locally with the RMA algorithm using specific Bioconductor packages. For data generated by other platforms, the normalized matrix was directly downloaded for analysis. Gene differential expression was assessed using LIMMA package from Bioconductor.

2.2. Hierarchical clustering

A hierarchical clustering algorithm was applied using gene’s log2 fold change (FC) from each analyzed experiment as input values. A distance matrix was computed among all the experiments within the database, using the Euclidean distance as a metric. The agglomeration method of the clustering process was set to complete. Heatmaps.2 library was used for dendrogram and heatmap generation. All the statistical computations were performed using R software. To evaluate the batch effects presence, normalized gene average expression data for each experiment was used as data input for hierarchical clustering analysis.

2.3. Functional enrichment

Genecodis3 software was used for functional enrichment using default parameters and selection of GO Biological Process as target annotations.

2.4. Statistical analysis

Moderated *t*-test statistics were applied to microarray features once a linear model was fitted. Statistical significance of the overrepresented GO biological processes in our target gene list was obtained with chi-square test. False discovery rate (FDR) method was employed to adjust the obtained p-values.

3. Results

3.1. Data collection

Experimental gene expression data corresponding to nutrigenomics experiments was identified from GEO database by launching specific queries. Results were filtered in order to obtain gene expression data from *Homo sapiens* as organism, and expression profiling by array as study type. Few of these studies, corresponding to human nutritional interventions with large cohorts, were filtered out since we were strictly interested on experiments performed on cultured cells. We initially identified 71 potential GEO studies (Table S1) to be included in our analysis. Of those, 34 studies were filtered out due to different criteria such as studies corresponding to human interventions, lack of replicates in the experimental designs, expression data obtained with rare or custom arrays, and expression data corresponding to micro RNA’s. We ended up with a set of 37 GEO studies.

3.2. Gene expression analysis workflow

Experiments included in each study were carefully assessed before analysis in accordance with their experimental design, by manually assigning control and perturbation samples. That is to say, for each experiment their appropriate control was obtained within the same study. Subsequently, a common computational analysis workflow applying linear models was used to assess differential expression in each experiment. Finally, microarray features were annotated with Gene Symbol and Entrez gene identifiers to allow cross-platform data integration. Thus, we obtained a database which includes gene differential expression data from 81 comparisons among different compounds, treatments and cell types (Fig. 1) that arise from the 37 GEO studies.

3.3. Cluster analysis

The clustering has been performed using log2 fold change (FC) expression values obtained following the gene expression analysis workflow. After removing missing values and aggregating expression values for duplicate gene ID's, we proceeded to integrate gene expression data from all the microarray platforms used in our database. Our database included expression data obtained from 19 distinct microarray platforms. A first limitation is that only overlapping genes represented in all platforms could be used in our analysis. Therefore, we used the corresponding Entrez gene ID of the features screened in each platform for data integration. Indeed, gene symbols can be hard to match across platforms because of the continuous updates of gene names, as well as the many to one relationship issues where different gene symbols might correspond to the same gene.

We ended up with a log2 FC expression matrix of 15,591 genes among 81 variables (experiment comparisons). Such an expression matrix included NA values corresponding to the genes that were absent in a microarray platform. With the aim of grouping experiments which trigger similar gene expression profiles across the studied cell lines and treatments, we performed a hierarchical clustering on the nutrigenomics gene expression matrix obtained from our database (Fig. 2).

We observed in the cluster dendrogram that the most remarkable property is that, as previously observed (Lamb et al., 2006), cell lines

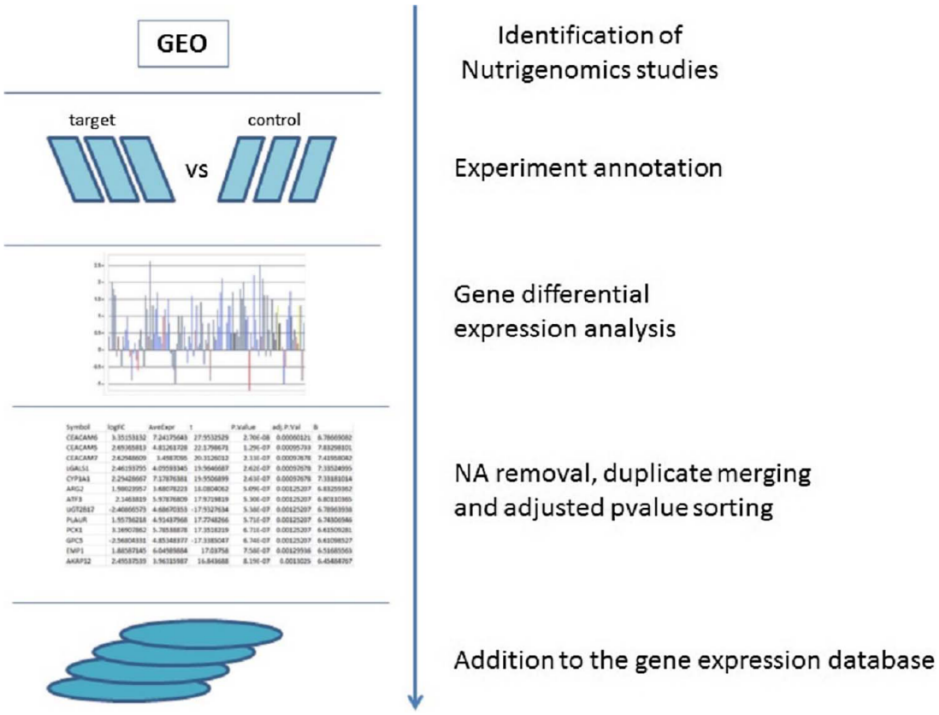


Fig. 1. Workflow used to set up our gene expression database from the identified nutrigenomics studies available in GEO database. The experiment annotation was done manually after carefully analyzing the corresponding experimental design. Gene differential expression analysis was performed using LIMMA models.

from common tissues tend to be clustered together. This is coherent and it is probably due to the expression of groups of genes which are cell specific. For example, we identified separate clusters grouping HeLa cells after treatment with different Tocotrienols (TCT) (alpha, delta and gamma), Myotubes treated with different fatty acids (eicosapentaenoic, linoleic and oleic acids), adipocytes CHUB-S7 cells treated with different concentrations of folate and vitamin B, and placental cells exposed to fish oil. Interestingly, clustering based on gene expression appears to be sensitive to the cell transitional state, as shown in the case of Keratinocytes treated with retinoic acid. Indeed, differentiating and proliferating Keratinocytes are found in separate clusters, independently of the retinoic acid type, either retinoic acid or 3,4-didehydroretinoic acid, used for treatments.

When paying attention to the clustering analysis from a compound point of view, interesting relationships are identified. There is a relatively big cluster (green) which groups different cell types after treatments with molecules like TCT (compounds naturally occurring in vegetable oils), lysophosphatidic acid (a natural bioactive lipid with growth factor-like functions), hydroxycholesterol, two amorfrutin (AMF) extracts and treatments with different strains of Lactobacillus (LB). Paradoxically, four different AMF extracts cluster in two relatively distant clusters. Those treatments were performed with the same concentration of 30 μM which evidences two distinct biological activities between AMF extracts 2–3 and AMF extracts 1–4.

Thus, we can infer that these compounds may share common biological properties in a cell type independent manner. Indeed, all of these compounds have previously demonstrated promising antioxidant and anti-inflammatory properties (Ahsan, Ahad, Iqbal, & Siddiqui, 2014; Contos, Ishii, & Chun, 2000; Fuhr, Rousseau, Plauth, Schroeder, & Sauer, 2015; Richard, Kefi, Barbe, Bausero, & Visioli, 2008). With the aim of getting insights into the effect of such compounds on cells from a functional point of view, we searched for common features among the top 3000 statistically significant differentially expressed genes in each experiment (green cluster, Fig. 2). However, we were not successful at identifying any significant overlap among the top 3000 statistically significant differentially expressed genes.

3.4. Treatments with potential anticancer natural compounds cluster together

One of the identified and clearly independent cluster (blue), groups together different types of cancer cell lines (MCF7, HT29 and subtypes of MDA cells) exposed to different bioactive compounds with previously reported anticancer properties, such as Indole-3-carbinol (I3C) (Quirit et al., 2017), a rosemary (RSM) extract, carnosic acid (CA) (Valdés, Sullini, Ibáñez, Cifuentes, & García-Cañas, 2015), withaferin A (WFNA) (Szarc vel Szic et al., 2014), sulforaphane (SFP) (Agyeman et al., 2012) and resveratrol (RVT) (Varoni, Lo Faro, Sharifi-Rad, & Iriti, 2016). We hypothesize that these compounds might be exerting their potential anticancer activity by a common mode of action on those cancer cells. 2 experiments corresponding to treatments with a RSM extract, at concentrations of 60 and 100 μg/mL respectively on SW620 cells, were not grouped within the anticancer cluster. This fact might be explained because the compound concentration used for this treatment was too high and triggered a pronounced cell death.

However, we should also consider that all of these clustering observations may be influenced by previously reported non-biological factors such as experiments generated in different laboratories/technicians, and experimental protocols (Luo et al., 2010), the so called batch-effects. Such a bias appears whenever merging expression data from different microarrays experiments for an integrative analysis, prior to a gene differential expression analysis, which should yield more robust statistics on differential gene expression results in most cases, for instance when studying gene expression levels in a specific disease or following medical treatment. In our case, each experiment has been analyzed independently and the resulting gene differential expression data (log2 FC expression values) has been combined in a large matrix in order to perform a data mining exercise. Because in this work we are not performing an integrative analysis, our results cannot be affected by such bias. To further test that batch effect does not influence our analysis, we performed a clustering analysis of a matrix containing normalized expression values from the same experiments (Fig. S1), which would mimic an integrative analysis prior to the gene differential expression processing. In this new dendrogram we can clearly appreciate the batch-effects bias. Indeed, all the included experiments strictly

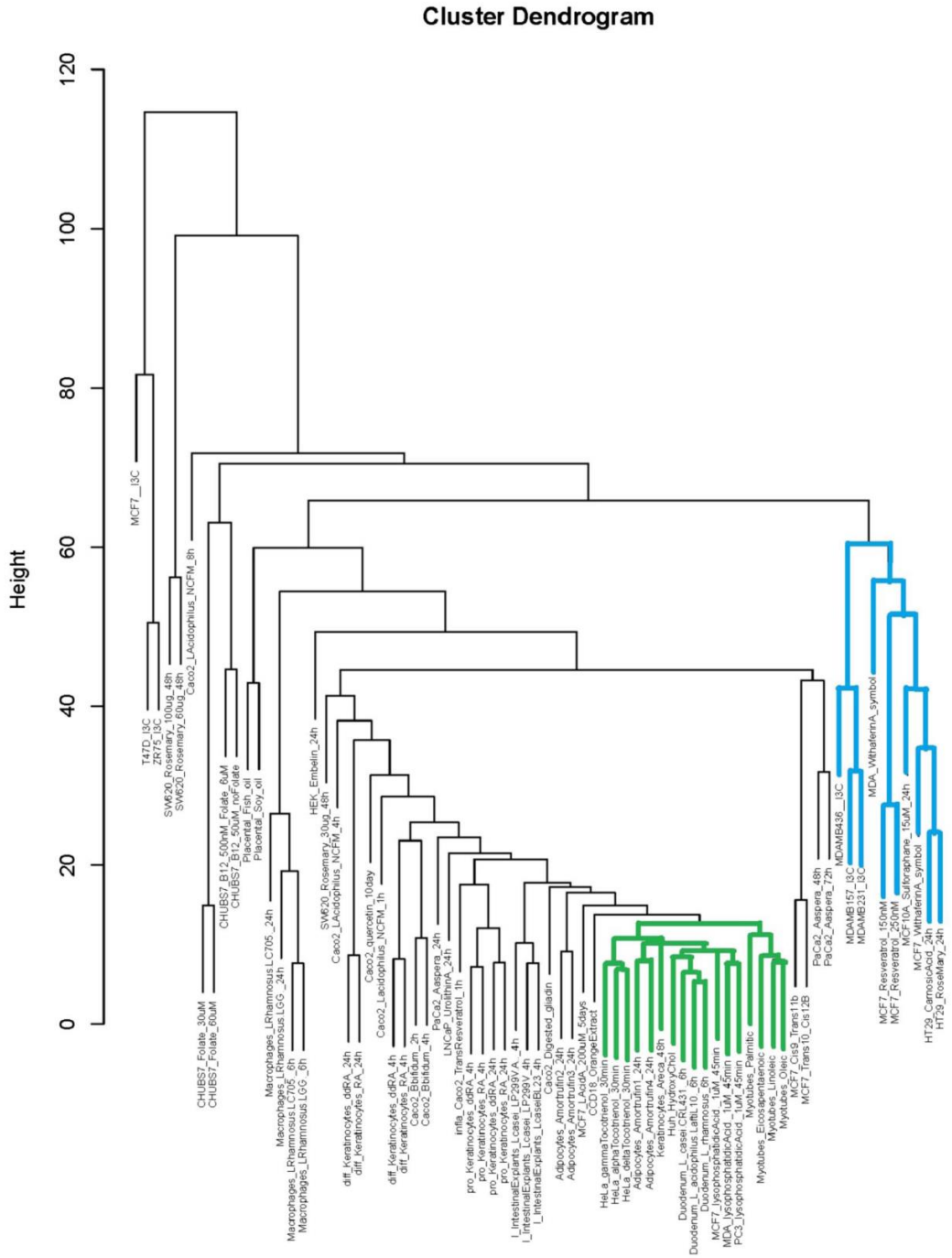


Fig. 2. Cluster dendrogram of the hierarchical clustering result. Log2 fold change values obtained after differential expression of each experiment were used as input data. The two clusters highlighted in blue and green color correspond to clusters grouping experiments with potential cancer protective and anti-oxidant/anti-inflammatory compounds respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cluster by study (experiments generated in the same lab), revealing the presence of batch-effects, and as a consequence no biological connections are detected anymore among the studied food compounds. Therefore, we can assume that our clustering analysis is not affected by the so called batch-effects.

We decided to compare the most statistically significant differentially expressed genes of each experiment in order to get insights into a common mode of action for these potential anticancer compounds. Therefore, in order to identify common and stable differential expressed genes, we chose one experiment from the most populated branch of this cluster as a representative for each different anticancer compound: CA, WFNA, SFP and RVT at 150 nM. As previously stated these 4 natural compounds have previously shown potential antiproliferative effects on cancer cells. Experiments corresponding to treatments with I3C were not included in the comparison since they were found in a separated branch of the same cluster. The RSM treatment performed on HT29 cells experiment was not included in this comparison since CA is mainly found in RSM. In addition, RSM and CA experiments show a high correlation since they are found in two close branches within the same cluster. As well as for RVT treatments, since both experiments are found in the same cluster, we decided to include the treatment with 150 nM as a representative for RVT treatments.

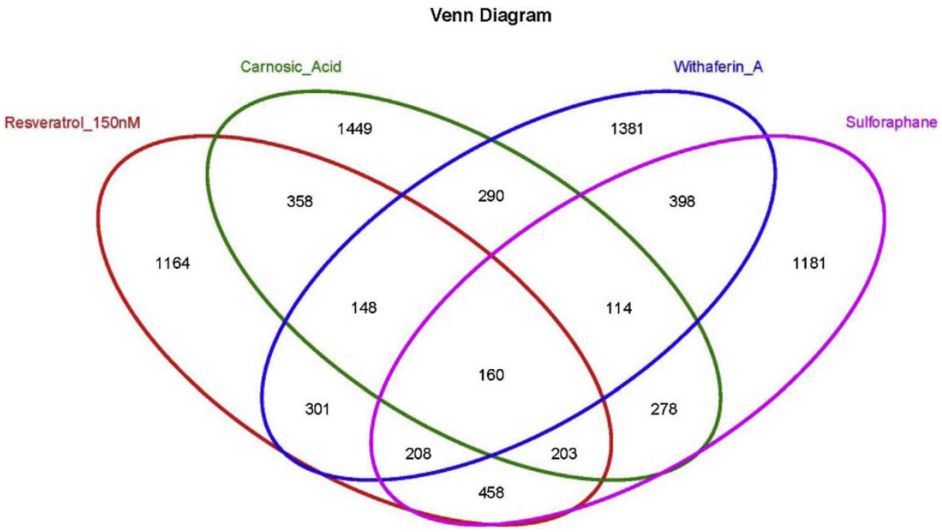


Fig. 3. Venn diagram showing the intersect among the top 3000 differentially expressed genes from four experiments found within the identified cancer protective cluster. Those potential cancer protective compounds are Resveratrol, Carnosic Acid, Withaferin A and Sulforaphane.

3.5. Identification of a cancer protective gene expression signature

Experiments from these 4 treatments were analyzed independently. We found that the top 3000 statistically significant differentially expressed genes shows an important overlap among these 4 experiments (Fig. 3). Therefore, we decided to focus on a common set of 160 differentially expressed genes.

A functional analysis was carried out in order to identify biological processes enriched in this set of differentially expressed genes (Table S2). The biological enrichment revealed the term GO:0,000,278 (mitotic cell cycle), grouping a set of 18 genes, as the most statistically significant item. This finding is very interesting and makes sense from an anticancer activity point of view.

Interestingly, analysis of the log2 FC expression values of these 18 genes among all the experiments included within our anticancer cluster shows a high correlation (Fig. S2) as well as a common pattern of down regulation among the ten experiments (Table S3). To further inspect this gene expression pattern across our nutrigenomic database, we generated a heatmap from the expression values of this set of genes (Fig. 4). As expected, the anticancer compounds within the larger branch of the previously identified anticancer cluster (WFNA, SFN, RVT, and RSM) are tightly grouped together based solely on the gene expression of this set of genes. Treatments with I3C on different cancer cell lines are also grouped with those compounds, but the level of downregulation of the genes included within the identified GES varies: the downregulation is stronger for I3C treatments on ZR75, MCF7 and T47D cell lines than for MDA cell line subtypes. The treatment with CA on HT29 is very close to the main cluster grouping the anticancer compounds, also due to the scarce downregulation level of the genes identified in our GES. Surprisingly, two additional experiments are closely related to the ones included in our anticancer cluster: Macrophages exposed to different strains of the probiotic LB rhamnosus, and CaCo2 cells exposed to transresveratrol (TRVT). Such observation might be explained because of the close connection between cell cycle and oxidative biological processes, and microorganisms as LB have previously shown positive effects on cellular oxidative stress (Forsyth et al., 2009). We also remark that, among the 3 concentrations used in the experiment of SW620 cells treated with RSM, only the treatment with a concentration of 100 μg/mL is grouped with those anticancer compounds. Thus, in this way we are able to confirm that higher concentrations of this compound are more effective to trigger cancer cell death as previously observed (Gonzalez-Vallinas et al., 2014). In overall, our results suggest that is possible to integrate publicly available data from different laboratories in order to find biological connections among experiments and thus to make new discoveries underlying the main hypothesis of the original experiment.

4. Discussion

Publicly available gene expression data allowed us to set up a local database in order to explore the gene expression landscape of cultured cell lines in response to different treatments with potentially bioactive compounds used in human diet. We performed a cluster analysis in order to classify those experiments based on gene expression data with the aim of getting insights into potential similarities of the effect of such compounds on those cells at gene expression level. The histological component revealed to be the primary identifiable factor dominating the clustering process. Additionally, the physiological attributes of cells, such as proliferation or differentiation, demonstrated to directly influence gene expression and hence the clustering process.

We identified experiments from tumor-derived cells lines treated with potential anticancer compounds showing strong similarities at gene expression level. After examination of these experiments independently, we found a GES of 18 genes which may characterize the protective effect against cancer shown by food compounds that can be found in broccoli sprouts (SFN) and plants from different families (I3C, WFNA, CA and RVT). Indeed, some of those have been traditionally used in ayurvedic medicine, as it is the case for WFNA. Further analysis of this set of genes across the whole database demonstrates that their expression level successfully groups together the experiments previously identified in our anticancer cluster.

This finding seems to be particularly significant, considering that gene expression data from the experiments included in our anticancer cluster was generated among 5 different laboratories and 4 different microarray platforms. Thus, we were able to identify stable biological relationships among experiments, overcoming potential batch effects which occur when integrating and analyzing data generated in different laboratories prior to differential gene expression analysis.

The genes included in this molecular signature (Table S4), which show a common downregulation trend after treatment with potential anticancer food compounds, appear to be mainly implied in mitotic and cell cycle processes. Such genes codify proteins required for DNA replication, chromosome alignment and segregation during mitosis (CENPE, CENPF, CENPK, GINS2, MCM3, KNTC1, PRIM1, RFC4, TYMS), microtubule assembly and organization (NDE1, TUBA4A, KIF2C), proteins involved in signal transduction through phosphorylation events (SKP2, CDK4, CDC7), a cyclin (CCNB1) and a nucleoporin (NUP107). Intriguingly, we also find in this GES a kinesin family member (KIF20A) with unknown function, for which a high expression level has been recently associated with poor prognosis in cancer disease (Liu et al.,

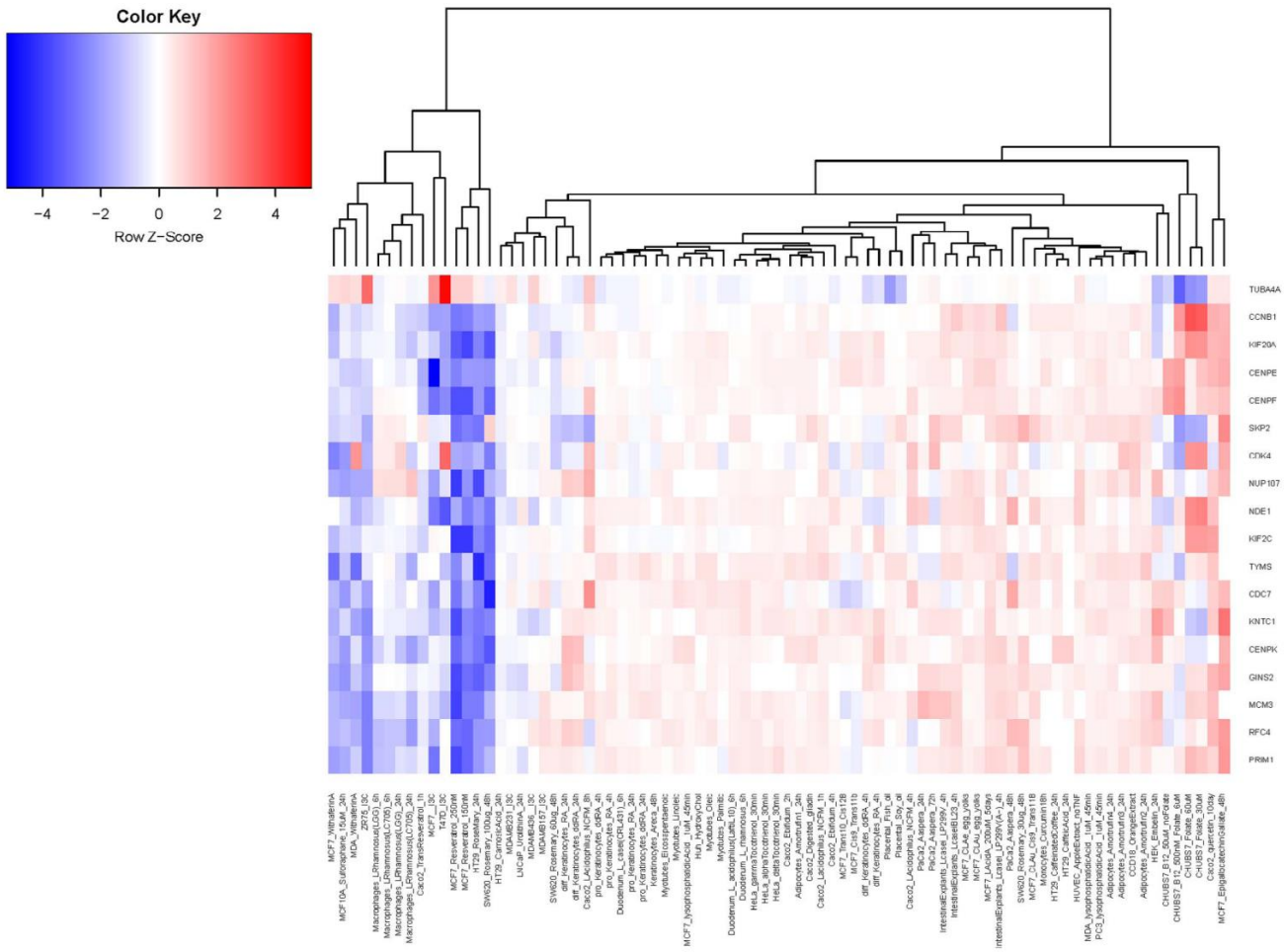


Fig. 4. Heatmap of nutrigenomics experiments based on the identified cancer protective gene signature. The represented log2 fold change values were normalized at row level.

2017; Zhang et al., 2016).

The present investigation shows the tremendous potential of publicly available experimental data for making new discoveries by applying in silico approaches. To our knowledge, this work is the first applying such data analysis approaches within the nutrigenomics field. Combining the results of multiple studies is a powerful tool that allows establishing solid correlations and relationships among variables, in our case gene expression, studied by different labs and platforms labs. In addition it also serves as a method for checking the reproducibility of results. We believe that in silico approaches are gaining weight in the present post-genomics era as more experimental data becomes publicly available. However the combination of such approaches with wet lab strategies, as gene expression confirmation by RT-PCR, will be a rule of thumb and would be hardly avoided. In our case, the original papers studying the effects of RVT, SFN, WFNA and CA focused on confirming the expression level of specific genes, which were found to be implicated in cholesterol metabolism and chromatin remodeling (Szarc vel Szic et al., 2014; Valdés et al., 2015).

Natural products have been an important source of lead compounds for drug discover, because of their vast chemical diversity, good drug-like properties and potential interactions with multiple cellular target proteins. For instance, the U.S. National Center for Complementary and Integrative Health (NCCIH) provides a list of a wide variety of natural products libraries (https://nccih.nih.gov/grants/naturalproducts/libraries) for screening purposes. The molecular signature identified in the data analysis presented in this work might provide potential biomarkers of cancer disease progression, and could therefore be used for prognosis. Importantly such a GES could be used as a benchmark for functional foods with potential cancer protective properties, tailored for human nutrition with a focus on both disease prevention and clinical nutrition purposes, thus setting up a framework for the use of functional food on human health.

Acknowledgements

This research was supported in part by the Spanish Agencia Estatal de Investigación and the European FEDER Funds (AGL2016-78922-R) to AD and RM-H. AD lab is supported in part by the Fundación Ramón Areces to the project “Modulation of exosomes transporting miRNAs and lncRNAs for intercellular communication as therapeutic tools in dyslipidemia treatment”.

Authors Contributions Statement

RM-H and AD designed the study. RM-H performed the study and analyzed data. RM-H and AD wrote the main manuscript text. All authors reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interest.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jff.2018.01.021.

References

Agyeman, A. S., Chaerkady, R., Shaw, P. G., Davidson, N. E., Visvanathan, K., Pandey, A., & Kensler, T. W. (2012). Transcriptomic and proteomic profiling of KEAP1 disrupted and sulforaphane-treated human breast epithelial cells reveals common expression profiles. *Breast Cancer Research and Treatment, 132*(1), 175–187.

Ahsan, H., Ahad, A., Iqbal, J., & Siddiqui, W. A. (2014). Pharmacological potential of tocotrienols: A review. *Nutrition & Metabolism (London), 11*(1), 52.

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular Pharmaceutics, 13*(7), 2524–2530.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Research, 41*(Database issue), D991–995.

Contos, J. J., Ishii, I., & Chun, J. (2000). Lysophosphatidic acid receptors. *Molecular Pharmacology, 58*(6), 1188–1196.

DeBusk, R. M., Fogarty, C. P., Ordovas, J. M., & Kornman, K. S. (2005). Nutritional genomics in practice: Where do we begin? *Journal of the American Dietetic Association, 105*(4), 589–598.

Ferguson, L. R. (2009). Nutrigenomics approaches to functional foods. *Journal of the American Dietetic Association, 109*(3), 452–458.

Forsyth, C. B., Farhadi, A., Jakate, S. M., Tang, Y., Shaikh, M., & Keshavarzian, A. (2009). Lactobacillus GG treatment ameliorates alcohol-induced intestinal oxidative stress, gut leakiness, and liver injury in a rat model of alcoholic steatohepatitis. *Alcohol, 43*(2), 163–172.

Fuhr, L., Rousseau, M., Plauth, A., Schroeder, F. C., & Sauer, S. (2015). Amorfrutins are natural PPARgamma agonists with potent anti-inflammatory properties. *Journal of Natural Products, 78*(5), 1160–1164.

Gil-Zamorano, J., Martin, R., Daimiel, L., Richardson, K., Giordano, E., Nicod, N., ... Davalos, A. (2014). Docosahexaenoic acid modulates the enterocyte Caco-2 cell expression of microRNAs involved in lipid metabolism. *Journal of Nutrition, 144*(5), 575–585.

Gonzalez-Vallinas, M., Molina, S., Vicente, G., de la Cueva, A., Vargas, T., Santoyo, S., ... Ramirez de Molina, A. (2013). Antitumor effect of 5-fluorouracil is enhanced by rosemary extract in both drug sensitive and resistant colon cancer cells. *Pharmacological Research, 72*, 61–68.

Gonzalez-Vallinas, M., Molina, S., Vicente, G., Zarza, V., Martin-Hernandez, R., Garcia-Risco, M. R., ... Ramirez de Molina, A. (2014). Expression of microRNA-15b and the glycosyltransferase GCNT3 correlates with antitumor efficacy of Rosemary diterpenes in colon and pancreatic cancer. *PLoS ONE, 9*(6), e98556.

Jia, Z., Liu, Y., Guan, N., Bo, X., Luo, Z., & Barnes, M. R. (2016). Cogena, a novel tool for co-expressed gene-set enrichment analysis, applied to drug repositioning and drug mode of action discovery. *BMC Genomics, 17*, 414.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., ... Golub, T. R. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science, 313*(5795), 1929–1935.

Lee, J. H., Kim, D. G., Bae, T. J., Rho, K., Kim, J. T., Lee, J. J., ... Kim, S. (2012). CDA: Combinatorial drug discovery using transcriptional response modules. *PLoS ONE, 7*(8), e42573.

Liu, S. L., Lin, H. X., Qiu, F., Zhang, W. J., Niu, C. H., Wen, W., ... Guo, L. (2017). Overexpression of Kinesin family member 20A correlates with disease progression and poor prognosis in human nasopharyngeal cancer: A retrospective analysis of 105 patients. *PLoS ONE, 12*(1), e0169280.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T., ... Zhang, J. (2010). A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics Journal, 10*(4), 278–291.

Pratanwanich, N., & Lio, P. (2014). Pathway-based Bayesian inference of drug-disease interactions. *Molecular BioSystems, 10*(6), 1538–1548.

Quirit, J. G., Lavrenov, S. N., Poindexter, K., Xu, J., Kyauk, C., Durkin, K. A., ... Firestone, G. L. (2017). Indole-3-carbinol (I3C) analogues are potent small molecule inhibitors of NEDD4-1 ubiquitin ligase activity that disrupt proliferation of human melanoma cells. *Biochemical Pharmacology, 127*, 13–27.

Ramirez de Molina, A., Vargas, T., Molina, S., Sanchez, J., Martinez-Romero, J., Gonzalez-Vallinas, M., ... Reglero, G. (2015). The ellagic acid derivative 4,4′-di-O-methylellagic acid efficiently inhibits colon cancer cell growth through a mechanism involving WNT16. *Journal of Pharmacology and Experimental Therapeutics, 353*(2), 433–444.

Richard, D., Kefi, K., Barbe, U., Bausero, P., & Visioli, F. (2008). Polyunsaturated fatty acids as antioxidants. *Pharmacological Research, 57*(6), 451–455.

Setoain, J., Franch, M., Martinez, M., Tabas-Madrid, D., Sorzano, C. O., Bakker, A., ... Pascual-Montano, A. (2015). NFFinder: An online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic Acids Research, 43*(W1), W193–199.

Snetselaar, L. (1994). Opportunities in the nutrition and food sciences: Research challenges and the next generation of investigators. *Journal of the American Dietetic Association, 94*(6), 598.

Szarc vel Szic, K., Op de Beeck, K., Ratman, D., Wouters, A., Beck, I. M., Declerck, K., ... Vanden Berghe, W. (2014). Pharmacological levels of Withaferin A (Withania somnifera) trigger clinically relevant anticancer effects specific to triple negative breast cancer cells. *PLoS ONE, 9*(2), e87850.

Valdés, A., Sullini, G., Ibáñez, E., Cifuentes, A., & García-Cañas, V. (2015). Rosemary polyphenols induce unfolded protein response and changes in cholesterol metabolism in colon cancer cells. *Journal of Functional Foods, 15*(Supplement C), 429–439.

Varoni, E. M., Lo Faro, A. F., Sharifi-Rad, J., & Iriti, M. (2016). Anticancer molecular mechanisms of resveratrol. *Frontiers in Nutrition, 3*, 8.

Zhang, W., He, W., Shi, Y., Gu, H., Li, M., Liu, Z., ... Zhang, Y. (2016). High expression of KIF20A is associated with poor overall survival and tumor progression in early-stage cervical squamous cell carcinoma. *PLoS ONE, 11*(12), e0167449.

Food & Function

PAPER

Cite this: *Food Funct.*, 2019, **10**, 4897

Identification and validation of common molecular targets of hydroxytyrosol†

María-Carmen López de las Hazas, ‡[a] Roberto Martin-Hernández, ‡[b] María Carmen Crespo,‡[c] João Tomé-Carneiro, [c] Lorena del Pozo-Acebo,[a] María B. Ruiz-Roso,[a] Joan C. Escola-Gil,[d,e,f] Jesús Osada, [g,h,i] Maria P. Portillo, [h,j] José Alfredo Martinez, [h,k,l,m] María A. Navarro, [g,h,i] Laura Rubió,[n] María José Motilva, [n,o] Francesco Visioli[c,p] and Alberto Dávalos *[a]

Hydroxytyrosol (HT) is involved in healthful activities and is beneficial to lipid metabolism. Many investigations focused on finding tissue-specific targets of HT through the use of different omics approaches such as transcriptomics and proteomics. However, it is not clear which (if any) of the potential molecular targets of HT reported in different studies are concurrently affected in various tissues. Following the bioinformatic analyses of publicly available data from a selection of *in vivo* studies involving HT-supplementation, we selected differentially expressed lipid metabolism-related genes and proteins common to more than one study, for validation in rodent liver samples from the entire selection. Four miRNAs (miR-802-5p, miR-423-3p, miR-30a-5p, and miR-146b-5p) responded to HT supplementation. Of note, miR-802-5p was commonly regulated in the liver and intestine. Our premise was that, in an organ crucial for lipid metabolism such as the liver, consistent modulation should be found for a specific target of HT even if different doses and duration of HT supplementation were used *in vivo*. Even though our results show inconsistency regarding differentially expressed lipid metabolism-related genes and proteins across studies, we found *Fgf21* and *Rora* as potential novel targets of HT. Omics approaches should be fine-tuned to better exploit the available databases.

Received 31st May 2019,
Accepted 13th July 2019

DOI: 10.1039/c9fo01159e

rsc.li/food-function

1. Introduction

3,4-Dihydroxyphenylethanol (hydroxytyrosol, HT), the main olive oil phenolic compound, is mostly found as part of complex (poly)phenols (secoiridoids),[1] which are easily hydrolyzed to yield HT after ingestion.[2] As digestion progresses, HT-derived metabolites (mainly its sulfate form) become the main compounds circulating in blood[3,4] and are recovered in urine.[5] The biological properties of HT have been widely investigated in different research areas including nutrition, medicine, pharmacology, chemistry and biotechnology.[6] This phenol is now considered as one of the most bioactive natural mole-

[a]*Laboratory of Epigenetics of Lipid Metabolism, Instituto Madrileño de Estudios Avanzados (IMDEA)-Alimentación, CEI UAM+CSIC, 28049 Madrid, Spain. E-mail: alberto.davalos@imdea.org; Tel: +34912796985*
[b]*Bioinformatics and Biostatistics Unit, IMDEA Food Institute, CEI UAM+CSIC, Madrid 28049, Spain*
[c]*Laboratory of Functional Foods, Instituto Madrileño de Estudios Avanzados (IMDEA)-Alimentación, CEI UAM+CSIC, 28049 Madrid, Spain*
[d]*Institut d'Investigacions Biomèdiques (IIB) Sant Pau, Barcelona, Spain*
[e]*CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain*
[f]*Departament de Bioquímica, Biología Molecular i Biomedicina, Universitat Autònoma de Barcelona, Spain*
[g]*Instituto Agroalimentario de Aragón, CITA-Universidad de Zaragoza, E-50013, Spain*
[h]*CIBER de Fisiopatología de la Obesidad y Nutrición, Instituto de Salud Carlos III, Madrid, E-28029, Spain*
[i]*Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Veterinaria, Instituto de Investigación Sanitaria de Aragón-Universidad de Zaragoza, Zaragoza, E-50013, Spain*
[j]*Obesity and Nutrition Group. Dpt. Nutrition and food Science, Faculty of Pharmacy, University of the Basque Country and Lucio Lascaray Research Center, Spain*
[k]*Department of Nutrition, Food Science and Physiology and Centre for Nutrition Research, University of Navarra, 31008 Pamplona, Spain*
[l]*Navarra Institute for Health Research (IdiSNA), 31008 Pamplona, Spain*
[m]*Precision Nutrition and Cardiometabolic Health Department, Madrid Institute of Advanced Studies (IMDEA Food), 28049 Madrid, Spain*
[n]*Food Technology Department, XaRTA-TPV, Escuela Técnica Superior de Ingeniería Agraria, University of Lleida, Av/Alcalde Rovira Roure 191, 25198 Lleida, Spain*
[o]*Instituto de Ciencias de la Vid y del Vino (ICVV) (Consejo Superior de Investigaciones Científicas, CSIC-Universidad de la Rioja-Gobierno de La Rioja), Logroño, Spain*
[p]*Department of Molecular Medicine, University of Padova, Viale G. Colombo 3, 35121 Padova, Italy*

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c9fo01159e

‡These authors contributed equally to this work.

cules.[7] *In vitro*, HT is an antioxidant[8] and its intake may beneficially influence cardiovascular disease (CVD) risk, *via* its potential to induce anti-atherosclerotic,[9] hypotensive, antioxidant, anti-inflammatory,[10] and hypocholesterolemic effects.[11] Of note, the European Food Safety Authority (EFSA) issued a health claim based on consistent results regarding the protective effects of olive polyphenols against the oxidation of blood lipids.[11]

Investigations that use cutting-edge, high-throughput, techniques referred to as "omics" are increasingly popular.[12] Omics tools have allowed the deepening of the knowledge on metabolism changes and identifying new potential disease (such as CVD) biomarkers, in a way that was not possible by genetic techniques alone.[13] In this sense, dietary intervention studies have successfully used transcriptomics and proteomics to show the mode through which diet induces alterations in gene and protein expression, providing information about the mechanisms of action and pathways regulated by micronutrients and helping in the identification of new biomarkers.[14] However, in contrast to other fields of study,[15] very few initiatives focusing on the establishment of databases that integrate largescale nutritional and genomics or genetics data have been developed.[15–17] The majority of such actions are aimed at standardizing nutritional studies and some publications – sometimes – over-emphasize the results obtained *via* database interrogation. Indeed, the scientific literature only describes two examples of such large-scale nutritional genomic data analysis: one related to functional genomics in chicken (Dhanasekaran *et al.*, 2014)[18] and the other one to genomic responses triggered by food bioactive compounds.[12] Indeed, it is complicated to extract the most relevant information from the large amount of data being produced worldwide, although such an approach would greatly strengthen scientific conclusions.[19]

By integrating transcriptomic and proteomic data, our initial goal was to identify consistently modulated potential molecular targets of HT reported in different studies where this compound was supplemented *in vivo*. We then used liver samples, a key tissue in lipid metabolism, obtained from different HT rodent studies to evaluate whether the previously identified candidates could be considered as solid targets of HT.

2. Materials and methods

2.1 Data collection and gene selection

From the PubMed and Scopus scientific databases, we gathered studies on *in vivo* supplementation with HT or its phenolic precursors, where gene and protein differential expressions were screened. Specific queries were launched with keywords such as "hydroxytyrosol AND proteomic", "hydroxytyrosol AND transcriptomic", "hydroxytyrosol AND gene", "hydroxytyrosol AND protein", "hydroxytyrosol AND miRNA", "hydroxytyrosol AND mRNA", "hydroxytyrosol AND genomic". We generated Venn diagrams of data from the selected studies, by means of in-house R scripts, to find intersections among differentially expressed genes.

Functional enrichment. Genecodis3 software was used for functional enrichment using default parameters and selecting GO biological processes as target annotations.

Statistical analysis. Moderated *t*-test statistics were applied to microarray features once a linear model was fitted. Statistical significance of the overrepresented GO biological processes in the target gene list was assessed through the chi-square test. The false discovery rate (FDR) method was employed to adjust the obtained *p*-values.

2.2 Ethics statements

All animal studies were approved by the respective Animal Ethics Committee of the institutions where each animal experimentation took place, namely: University Complutense of Madrid (CEA-UCM 93/2012; CEEA 10-06/14, 31st July 2014), University of Lleida, and Universidad Mixta de Investigación, Zaragoza. All procedures followed the *Guide for the Care and Use of Laboratory Animals*, published by the US National Research Council (Eighth Edition, 2010), except for the study by Acin *et al.* (Study 3, see below), which is prior to 2010 and followed the Ethical Committee for Animal research of the University of Zaragoza.

2.3 Brain and liver microarray analysis

Gene expression profiles in brain and liver tissues were analyzed using the Illumina MouseRef-8 v2 Expression BeadChip® platform with Ambion Labelling. Four biological replicates per group were included. This BeadChip targets approximately 25 600 well-annotated RefSeq transcripts, representing over 19 100 unique genes. Data were background-corrected and normalized using GenomeStudioTM Software (Illumina, San Diego, CA, USA) and following the manufacturer's instruction. Differential expression was assessed using the Limma's Bioconductor package in the R statistical programming environment.

2.4 Liver samples used for transcriptomic and proteomic validations

Transcriptomic and proteomic validations were performed in liver samples obtained from a selection of four previously published studies where HT was administered to rodents, as follows: (1) study from Tomé-Carneiro *et al.* (2016), hereon referred to as Study 1. Briefly, in this study young C57BL/6 mice (2 months old, n = 14) were fed a purified control diet alone (n = 7) or supplemented with approximately 45 mg HT per kg bw per day (Seprox Biotech, Madrid, Spain) (n = 7), for 8 weeks.[20] (2) A second cohort from the Tomé-Carneiro *et al.* (2016)[20] study, hereon referred to as Study 2. Briefly, in this acute ingestion study, 15 mg of HT dissolved in water (Seprox Biotech, Madrid, Spain) were administered (by gavage) to young C57BL/6 mice (10 weeks old), which were sacrificed immediately (control group, n = 9) or 4 h after ingestion (n = 9). (3) Study from Acín *et al.* (2006), hereon referred to as Study 3. Briefly, for 10 weeks, 14 homozygous apoE KO mice

(2 months old, n = 14) were given free access to drinking water (control group, n = 7) or to an aqueous solution and provided a dose of 10 mg HT per kg per day (n = 7).[21] (4) Study from Catalán *et al.* (2016), hereon referred to as Study 4. Briefly, female Wistar rats (300–350 g, n = 8) were fed a standard diet (SD) (control group, n = 4) or SD supplemented with 5 mg of secoiridoids per kg per day (n = 4), for 21 days.[22]

2.5 Transcriptomic validations

Commonly differentially expressed genes identified in the Venn diagram (Fig. 1) were used for transcriptomic validations. Oligonucleotide primers were designed to amplify the selected genes in both mouse and rat species. Gene function and primers are listed in Table 1. Briefly, total RNA was isolated from liver samples using a Trizol/Qiagen RNeasy kit. cDNA synthesis was performed with 1 μg of RNA using the miScript II RT kit (Qiagen) according to the supplier's instructions. qPCR was carried out in a ABI 7900 HT Real-time PCR system with a 384 well plate format, using the FastStart Essential DNA Green Master mix (Roche, Switzerland) at 95 °C for 10 min, followed by 40 cycles at 95 °C for 15 s and 58 °C for 1 min. Gene expression was normalized with respect to *Gapdh* expression, and relative quantification was calculated using the $2^{\Delta\Delta Ct}$ method.

2.6 Proteomic validations

For proteomic validations, protein selection was based on the commonly differentially expressed proteins identified in the corresponding Venn diagram (Fig. 3). Antibodies used and functions of the selected proteins are described in Table 2. Briefly, the liver samples from the four above-mentioned studies were homogenized in RIPA buffer (200 mM sodium orthovanadate, 1 mM Pefabloc SC, and 2 mg mL^{-1} protease inhibitor cocktail (Sigma, Madrid, Spain)), sonicated for 5 min and frozen overnight at −80 °C. After centrifugation at 12 000g for 30 min at 4 °C, the supernatant was collected for total protein quantification using a Pierce BCA Protein Assay kit (Thermo Fisher Scientific) according to the manufacturer's protocol. For western blotting, 30 μg of total protein were electrophoresed in 6–15% SDS-PAGE gel and then transferred onto nitrocellulose membranes. The membranes were blocked with LiCor blocking buffer at RT for 1 h. Overnight incubation at 4 °C was performed with a primary antibody followed by incubation (1 h, RT) with the LiCor fluorescent secondary antibody (goat anti-Mouse IRDye® 800 CW or goat anti-Rabbit IRDye® 680 CW). The membranes were visualized using a LiCor Odyssey scanner. Proteins were quantified by densitometry using Image Studio Lite 5.2.5 analytical software (LiCor, Lincoln, NE) and normalized to GAPDH.

2.7 miRNA analysis

For miRNA analysis, an unbiased whole genome miRNA analysis was performed in the mice liver samples (n = 5 per group) from Study 1 using small RNA sequencing. RNA integrity was determined using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Following the manufacturer's protocol, a NEBNext® multiplex small RNA Library Prep Set for Illumina (New England BioLabs, Ipswich, MA) was used to prepare the libraries, and sequenced using the Illumina NextSeq 500 platform. After trimming adapter sequences, Bowtie2 was used for read alignment against high confidence mouse mature miRNA sequences, obtained from the miRBase

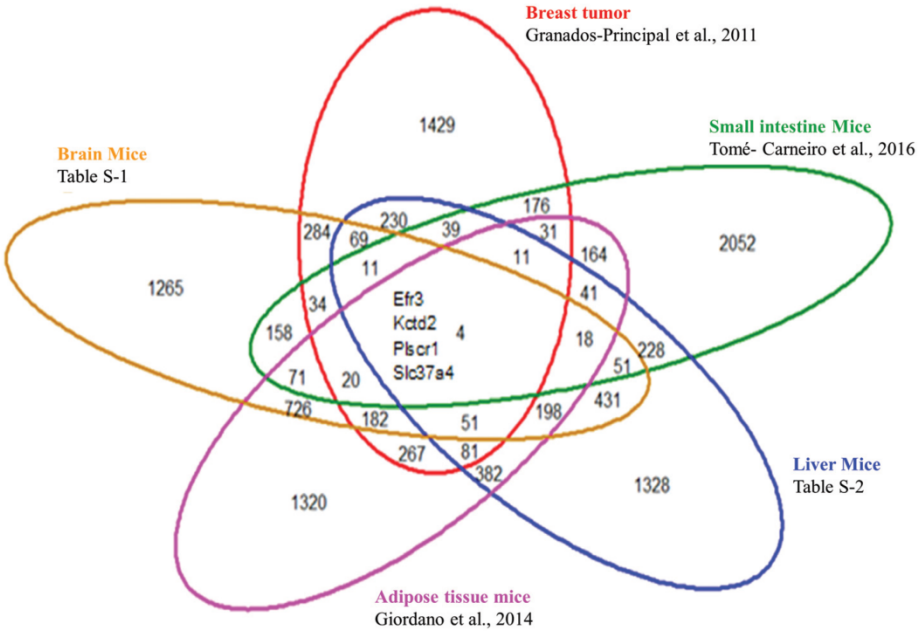


Fig. 1 Venn diagram showing the common differentially expressed genes identified through bioinformatic analysis. Transcriptomic data available in public repositories from *in vivo* dietary supplementation with hydroxytyrosol. Gene differential expression was analyzed using LIMMA models.

Table 1 Function of the genes and list of primers designed for the validation of the studies

Gene	Primers		Function
	Forward	Reverse	
B-*Efr3a*	CTTTGCGTCCTCGCTACAAAC	CCATATCAGCTTTAACAAGGCCA	Involved in the functional maintenance of sensory and motor nervous tissues.
B-*Kctd2*	CCTACTTCGTGACCACCAGAC	GAGTTTTCCATGGCGGAGGT	Potassium channel tetramerization domain containing protein 2.
B-*Plscr1*	GGTCCGTGTGTTGTGTGTAG	TGCTCCTCGTTTCCAGTTCTT	Involved in phosphatidylserine externalization regulation during cell activation.
B-*Slc37a4*	AACCGCAAAACCTTCTCCTT	TACGTTGACCAGACCAACCA	Regulates glucose-6-phosphate transport and maintains glucose homeostasis.
B-*Ppp1cb*	CAGAAGTCCGAGGGTTGTGTA	CAGATGGTTTCCAAAGACTGCTT	Involved in the regulation of cell division, glycogen metabolism, and muscle contractility.
B-*Tjp2*	GTTTGCCGTTCAGCAGCTTAG	CTTCAAAACCTCGGTCGTCAT	Component of the tight junction barrier in epithelial and endothelial cells.
B-Top1	GCCAAGGTGTTCCGTACCTA	TCAGGTCCTTTCGAGCATCT	Essential for cell growth and division *in vivo*.
B-Snx16	CCAGAAGAAAGCTGGGTAGTTTT	GGAAGTGCTAATCGAAAGCCTG	Involved in cholesterol transport, and transport of tetraspanin CD81.
B-Anks6	GGAGCTGGGGATTAAGACGG	TAGAATCTGCCTCTCACGCC	Plays a role in renal and cardiovascular development.
B-Plscr2	CTGGGTATGCCCCTCAGTATC	GGGAACTTGGTAGTTAGTCTGGA	Plays an active role in altering lipid asymmetry at the plasma membrane.
B-Soat1	GAAGGCTCACTCATTTGTCAGA	GTCTCGGTAAATAAGTGTAGGCG	Catalyzes the formation of fatty acid-cholesterol esters.
B-Fgf21	CAGATGTGGGTTCCTCCGAC	AAGATGCATAGCTGGGGCTT	Secreted endocrine factor that functions as a major metabolic regulator.
R-*Crot*	AAGCCGGGTGCAGGAGTTTT	CCACTCTTCCAGCCAGTTTCT	Plays a role in lipid metabolism and fatty acid beta-oxidation.
M-*Crot*	GAACGGACATTTCAGTACCAGG	CTTCATTTGCGAATGGTTTCACT	
B-Rora	GTGGAGACAAATCGTCAGGAAT	GACATCCGACCAAACTTGACA	Controls lipid homeostasis by negatively regulating the transcriptional activity of PPARγ, that mediates hepatic lipid metabolism.
B-Sorl1	CCCAGCCTATCCAGGTGTATG	CGGGCTAATGCCACGATCA	Binds LDL and transports it into cells by endocytosis.
B-Elovl1	GAAGAAGGACGGGCAAGTGA	TTGCAGCTGGGCATGAAGTA	Involved as the precursors of membrane lipids and lipid mediators.
B-Acsl4	CTCACCATTATATTGCTGCCTGT	TCTCTTTGCCATAGCGTTTTTCT	Plays a key role in lipid biosynthesis and fatty acid degradation.
B-Lipe	GTTACCACCCTGCAGTCCTC	AAGTGTCTCTCTGCACCAGC	Converts cholesteryl esters to free cholesterol for steroid hormone production.
B-Lpin2	GAAGTGGCGGCTCTCTATTTC	AGAGGGTTACATCAGGCAAGT	Plays a role in triglyceride metabolism.

Efr3a: EFR3 homolog A; *Kctd2*: potassium channel tetramerization domain containing 2; *Plscr1*: phospholipid scramblase 1; *Slc37a4*: solute carrier family 37 member 4; *Ppp1cb*: protein phosphatase 1 catalytic subunit beta; *Tjp2*: tight junction protein 2; *Top1*: DNA topoisomerase I; *Snx16*: sorting nexin 16; *Anks6*: ankyrin repeat and sterile alpha motif domain containing 6; *Plscr2*: phospholipid scramblase 2; *Soat1*: sterol O-acyltransferase 1; *Fgf21*: fibroblast growth factor 21; *Crot*: carnitine O-octanoyltransferase; *Rora*: RAR-related orphan receptor alpha; *Sorl1*: sortilin related receptor 1; *Elovl1*: ELOVL fatty acid elongase 1; *Acsl4*: acyl-CoA synthetase long chain family member 4; *Lipe*: hormone sensitive type lipase E; *Lpin2*: lipin 2; B: Primer designed for both species, *Mus musculus* and *Rattus norvegicus*; R: *Rattus norvegicus*; M: *Mus musculus*.

database. Finally, only reads showing a unique valid alignment against the reference sequences were considered for mature miRNA counting.

2.8. miRNA bioinformatic analysis

MicroRNAs' targets presenting hits on the 3′UTR position and showing a binding *P*-value score equal to 1 were obtained from the miRWalk 3.0 database[23] and used for further analysis. A functional enrichment of these genes, targeted by at least two of the differentially expressed microRNAs, was performed in the Panther database v.11[24] using Gene Ontology (GO) and Panther pathway annotations. A subset of four significantly modulated miRNAs in response to HT supplementation were used for Gene Interaction (GI) analysis using the above-mentioned target genes. GI analysis was performed as previously described[20] including target genes targeted at least by two miRNAs. The target dot size was directly correlated with the number of interactions with the set of miRNAs. As for the functional analysis, only the genes targeted simultaneously by at least two miRNAs are shown.

2.9. Statistical analysis

Comparisons between HT-supplemented groups and controls were performed by means of two-tailed *t* tests or Mann–Whitney tests; when assumptions for parametric testing were not met GraphPad Prism 7.02 (La Jolla, CA) was used. In all cases, $p < 0.05$ was considered as statistically significant.

3 Results

3.1 Identification of common differentially expressed genes

Specific searches in public scientific databases for *in vivo* interventions involving hydroxytyrosol supplementation returned scarce results (Table 3). The GEO database, which contains high throughput genomic and proteomic data among others,

Table 2 Selected proteins and types of antibodies used

Proteins	Company	Molecular weight (kDa)	Host	Function
CAR3	Thermo Fisher	29.6	Rabbit	Involved in oxidative stress.
FASN	Cell Signaling	273	Rabbit	Main function is to catalyze the synthesis of palmitate from acetyl-CoA and malonyl-CoA.
PRDX1	Cell Signaling	21	Rabbit	Belongs to a family of antioxidant enzymes. Reduction of hydrogen peroxide and alkyl hydroperoxides.
VIM	Cell Signaling	57	Rabbit	Involved in neurogenesis and cholesterol transport.
GAPDH	Sigma	37	Mouse	Housekeeping protein.
HSPD1	Bethyl	60	Rabbit	Involved in stress response.
ACTN4	Bethyl	110	Rabbit	Transcriptional coactivator, stimulating transcription mediated by the nuclear hormone receptors PPARG and RARA.

CAR3: Carbonic anhydrase 3; FASN: fatty acid synthase; PRDX1: peroxiredoxin 1; VIM: vimentin; HSPD1: heat shock protein family D (Hsp60) member 1; ACTN4: actinin alpha 4; GAPDH: glyceraldehyde-3-phosphate dehydrogenase.

Table 3 *In vivo* studies involving supplementation with hydroxytyrosol where transcriptomic analyses were performed

Model	Dose & time	Analysis	Aim of study	Reference
Male C57BL/6J mice	5 mg per kg bw HT.	qRT-PCR	Evaluate the molecular adaptations in the liver involved in the anti-lipogenic, anti-inflammatory, and anti-oxidant effects of HT	8
Male C57BL/6J mice	20 mg HT per kg bw per day 21 days.	qRT-PCR	Identify early, predictive biomarkers for WAT expansion.	70
Male db/db mice	10 or 50 mg HT per kg per day. 8 weeks	qRT-PCR	Evaluate the neuroprotective effects of HT in db/db mice and SH-SY-5Y neuroblastoma cells.	71
Sprague–Dawley rats	10 or 50 mg HT per kg per day. During gestation	qRT-PCR	Investigate the HT effect on the prenatal stress	72
Male C57BL/6 mice	~45 mg HT per kg bw per day. 8 weeks	Microarray qRT-PCR	Nutrigenomic effects of HT with specific reference to the adipose tissue and glutathione metabolism.	26
Female Sprague–Dawley rats	0.5 mg kg^{-1} 6 week	Microarray qRT-PCR	Hydroxytyrosol inhibits growth and cell proliferation and promotes high expression of sfrp4 in rat mammary tumours.	25
C57BL/6 male mice	0.03 g% HT 8 weeks	Microarray qRT-PCR	Chronic hydroxytyrosol feeding modulates glutathione-mediated oxido-reduction pathways in adipose tissue: a nutrigenomic study	26
Male C57BL/6 mice	0.03 g% HT 8 weeks	Microarray qRT-PCR	Hydroxytyrosol supplementation modulates the expression of miRNAs in rodents and in humans.	20

HT: Hydroxytyrosol.

only accounted for two studies concerning gene expression in humans (GSE75027 and GSE75026) after olive oil intake (where HT is preeminent). Moreover, only one study concerning an *in vivo* intervention with HT was found, where a breast tumor-induced model of *Rattus norvegicus* was used (GSE15944).[25] Finally, two sets of data from a study carried out in our laboratory (ESI Tables S1 and S2;† not available in scientific databases), regarding microarray screening in brain and liver tissues from diet-HT supplemented mice,[26] were included in this study. The aforementioned five sets of data were used to find mutual differentially expressed genes (Fig. 1). *Efr3*, *Kctd2*, *Plscr1* and *Scl37a4* were identified as differentially expressed after HT supplementation in all analyzed data sets.

3.2 Transcriptomic validations

After the identification of differentially expressed genes shared by at least three studies (Fig. 1 and ESI Table S3†), 18 genes associated with lipid metabolism were finally selected for validation. Validation was performed in the liver samples from HT-supplemented animal models *vs.* controls (Fig. 2) taken from the four selected studies (Studies 1–4, see Materials and methods for details). In the liver samples from Study 1 a statistically significant increase was seen in *Tjp2*, *Top1*, *Rora*, *Sorl1*, *Fgf21* and *Ppp1cb* (Fig. 2A). In the liver samples from Study 2 a statistically significant increase was seen in *Slc37a4*, *Snx16*, *Rora* and *Fgf21*, whereas a statistically significant decrease was observed in *Sorl1* (Fig. 2B). In the liver samples from Study 3 a statistically significant increase was seen in *Slc37a4*, *Anks6*, *Plscr2* and *Lipe* (Fig. 2C). In the liver samples from Study 4 a statistically significant increase was seen in *Sorl1* (Fig. 2D).

3.3 Identification of common differentially expressed proteins

Publicly available large-scale proteomics data regarding HT supplementation are scarce. To determine whether HT consumption affects specific signaling pathways, we comprehensively analyzed publications involving *in vivo* HT supplementation to extract protein expression information from the reported data tables (Table 4). Then, the collected proteomic data were subjected bioinformatic analyses to identify differentially expressed proteins common to at least two of these studies (Fig. 3). Bioinformatic analysis showed that ALDH2, SELENBP1, HSPD1, PPIA, VIM, YWHAG, RPL8, ACTN4, NPM1,

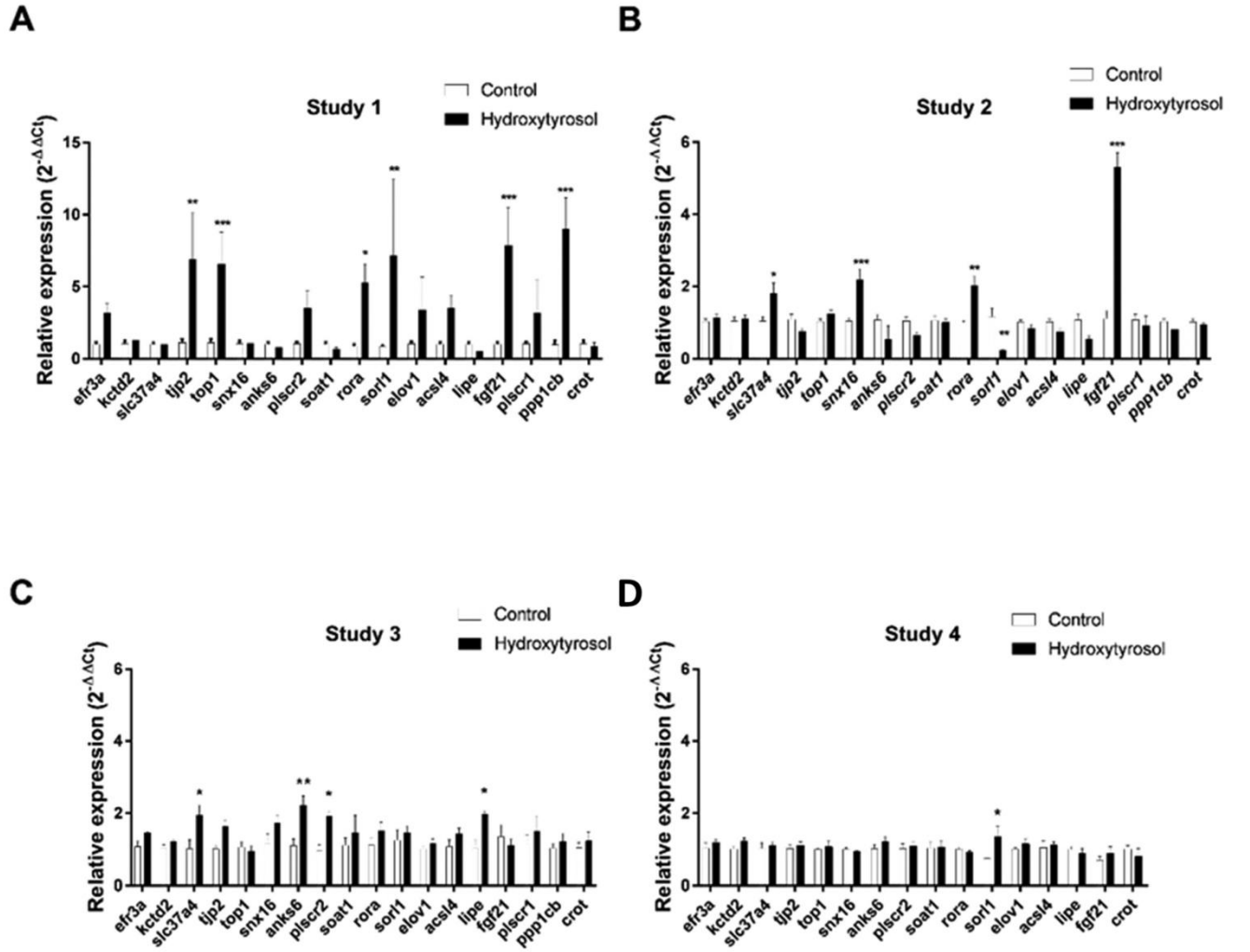


Fig. 2 Validation of common transcripts predicted to be modulated by hydroxytyrosol supplementation. A set of transcripts were chosen from bioinformatic analysis and validated in the liver samples of different intervention studies. Gene expression was analyzed by RT-qPCR. (A) Male young C57BL/6 mice (n = 7 per group) fed with a control or HT diet (45 mg HT per kg bw per day), for 8 weeks (Study 1). (B) Male young C57BL/6 mice were administered (gavage) an acute dose of 15 mg of HT (dissolved in water) and sacrificed 4 h after ingestion (n = 9) (Study 2). (C) Male young homozygous apoE KO mice (n = 7 per group) fed with an aqueous solution of 10 mg HT per kg per day (n = 7), for 10 weeks (Study 3). (D) Female Wistar rats (300–350 g, n = 4 per group) fed with a standard diet (SD) or SD supplemented with 5 mg HT per kg per day, for 21 days (Study 4). HT, hydroxytyrosol.

Table 4 *In vivo* studies involving supplementation with hydroxytyrosol where high throughput proteomic analyses were performed

Model	Dose & time	Aim of study	Ref.
Male C57BL/6 mice	~45 mg HT per kg bw per day. 8 weeks	Impact of long-term HT supplementation on the proteome in metabolically active tissues (adipose tissue and liver)	42
Female Wistar rats	5 mg kg^{-1} day^{-1} 21 days	Proteomic analyses in cardiovascular tissues (aorta and heart)	22
Male Rowett Hooded Lister rats	10 mg kg^{-1} diet. 12 weeks	Effects in the liver through proteomics and network analysis	43

HT: Hydroxytyrosol; ACTN4: alpha-actinin-4; RPL8: 60S ribosomal protein L8; ALDH2: mitochondrial aldehyde dehydrogenase.

ALB, HSP90AB1, CAR3, FASN, HBB-B1, PRDX1, CFL1, GLUD1, VCL, DLAT, and GPD1 proteins were common to at least 2 studies. Only two proteins, Hspd1 and Actn4, were common to three studies.

3.4 Proteomic validations

After the identification of differentially expressed proteins, common to at least two studies, validation was carried out in

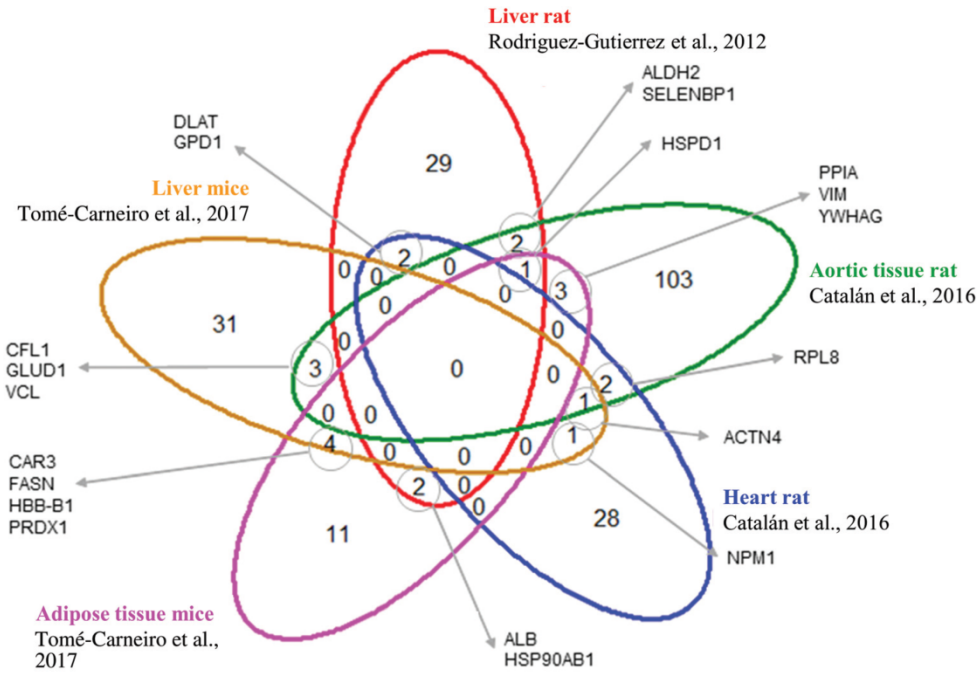


Fig. 3 Venn diagram showing the common differentially expressed proteins identified by means of bioinformatic analysis. Proteomic data available in the scientific literature from *in vivo* dietary supplementation with hydroxytyrosol.

the liver samples from the HT-supplemented animal models and controls (Studies 1–4). Overall, we did not find significant differences in any of the proteins analyzed in the livers of the HT- (or their secoiridoids precursors) supplemented animals compared with those of the controls (Fig. 4). In the liver samples from Study 1, the HT-supplemented groups showed a slight, statistically non-significant, decrease in VIM and increase in HSPD1, compared with those of the controls (Fig. 4A). In samples from Study 2, involving an acute ingestion of HT, a decrease in the expression was seen for PRDX1 and CAR3, and an increase in FASN, although statistical significance was not reached (Fig. 4B). As for Study 3, in HT-supplemented ApoE KO mice, non-significant decreases in the expression of ACTN4 and an increase in VIM were observed (Fig. 4C). Finally, a non-significant decrease in FASN was observed in the liver of female Wistar rats (Study 4) (Fig. 4D). CAR3 was not analyzed in Study 4 samples as the anti-mouse antibody used had no cross-reactivity to rats.

3.5 Post-transcriptional regulation by miRNAs

Post-transcriptional regulation is commonplace in biological systems and miRNAs bind complementarily to the 3′UTR sequence mediating negative post-transcriptional regulation,[27] in turn impacting the proteome.[28] Thus, we next assessed the modulation of miRNA levels and explored the potential impact it could have on the proteome. The liver samples from Study 1 were subjected to small RNA sequencing and miRNAs were analyzed (Fig. 5). From the 247 miRNAs detected in the mouse liver samples (ESI Table S4†), only 4 were found to be differentially expressed in the HT supplemented group after FDR adjustment (Fig. 5A). From these, miR-802-5p, miR-30a-5p and miR-146b-5p were up-regulated, whereas miR-423-3p was down-regulated. Because one gene can be regulated by different miRNAs, we also searched for validated targets likely to be modulated by more than one miRNA responding to HT treatment (Fig. 5B). Gene Interaction (GI) analysis was performed (see Materials and methods for details) generating a unique list of 279 genes potentially modulated by at least two miRNAs. The genes modulated by the four miRNAs included *Ccdc117*, *Ntrk2*, *Mrpl17*, *Timm22*, *Zfp945*, *Ubxn7*, *Tmem71*, *Slc30a7*, *Gucy1a2*, *4931406C07Rik*, *Zdhhc21*, and *Dclk1* (Fig. 5B). In particular, *Gucy1a2* is involved in an endothelin signaling pathway, *Zdhhc21* in metabolic processes (palmitoyltransferase), and *Timm22*/*Ntrk2* in cellular component organization. Gene ontology analysis of modulated miRNAs targets (by more than one miRNA) suggested their involvement in the regulation of major pathways, including the Wnt signaling pathway (P00057), the CCKR signaling map (P06959) or the inflammation mediated by the chemokine and cytokine signaling pathway (P00031), among others (Fig. 5C). Finally, none of the genes matched the ones obtained after the analysis of the transcriptomic or proteomic data sets, suggesting that their levels are not directly controlled by these specific miRNAs.

4 Discussion

Many nutritional intervention studies[29,30] demonstrated that food and its bioactive components affect the expression of genes, which can impact disease prevention.[31–34] High

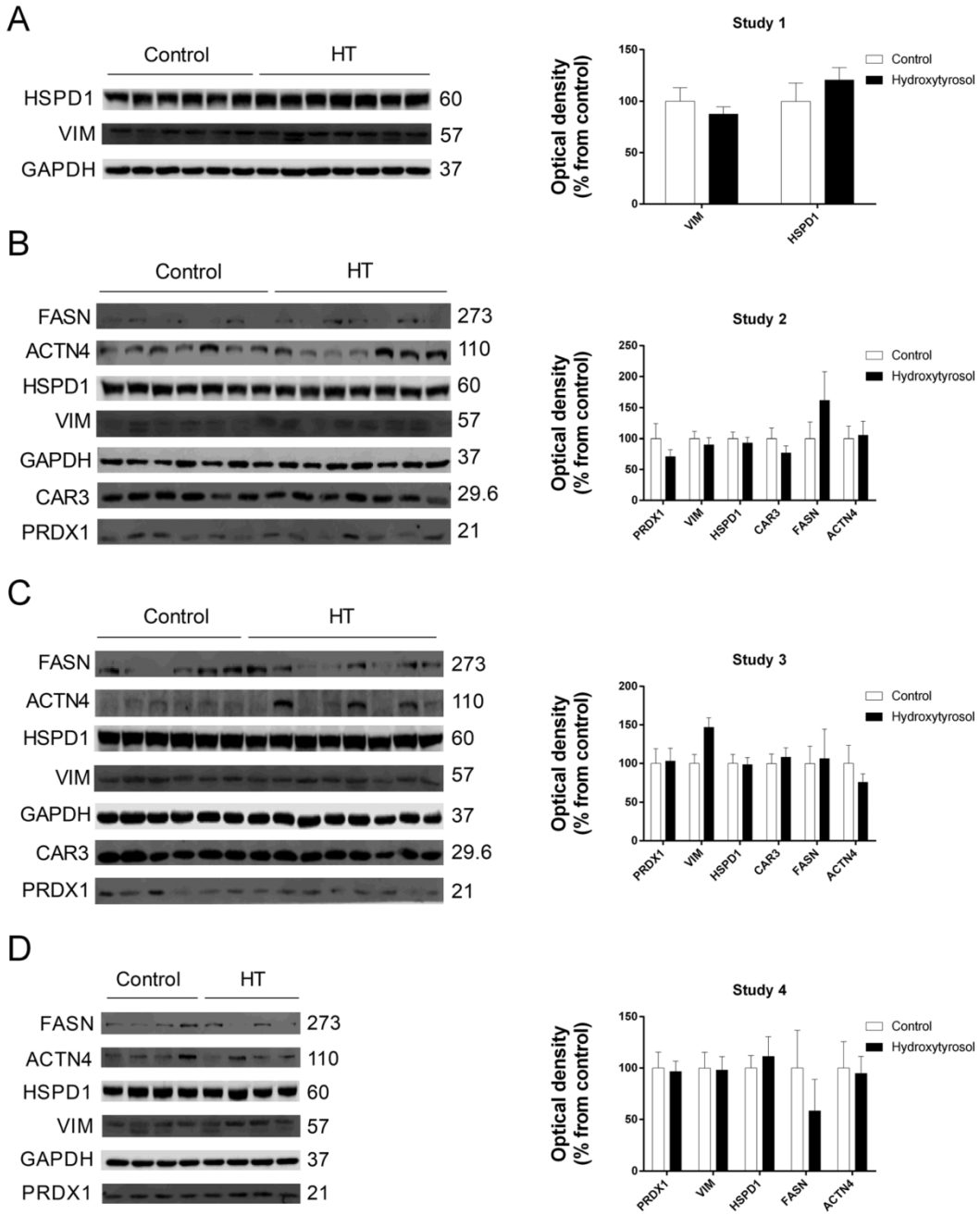


Fig. 4 Validation of common proteins predicted to be modulated by HT supplementation. A set of proteins were chosen after bioinformatic analysis of proteomic data and validated in the liver samples of different intervention studies. Protein expression was analyzed by western blotting. (A) Male young C57BL/6 mice (n = 7 per group) fed with a control or HT diet (45 mg HT per kg bw per day), for 8 weeks (Study 1). (B) Male young C57BL/6 mice administered (gavage) with an acute dose of 15 mg of HT dissolved in water and sacrificed 4 h after ingestion (n = 9) (Study 2). (C) Male young homozygous apoE KO mice (n = 7 per group) fed with an aqueous solution of 10 mg HT per kg per day (n = 7), for 10 weeks (Study 3). (D) Female Wistar rats (300–350 g, n = 4 per group) fed with a standard diet (SD) or SD supplemented with 5 mg HT per kg per day, for 21 days (Study 4). HT, hydroxytyrosol.

throughput transcriptome and proteome analysis can be very useful in the discovery of new biomarkers and pathways implicated in metabolic diseases.[35] Moreover, high throughput analyses aid in assessing the physiological effect that bioactive compounds exert on a wide variety of diseases such as diabetes,[36,37] obesity,[38] and cancer.[39,40] Hence, these techniques are useful to explore the mechanisms of action of nutrients and phytochemicals. Omics technologies are widely adopted to concomitantly study the expression of thousands of genes and proteins, generating a vast amount of data that accumulates over time and is generally available in public repositories. These data sets could potentially be exploited to

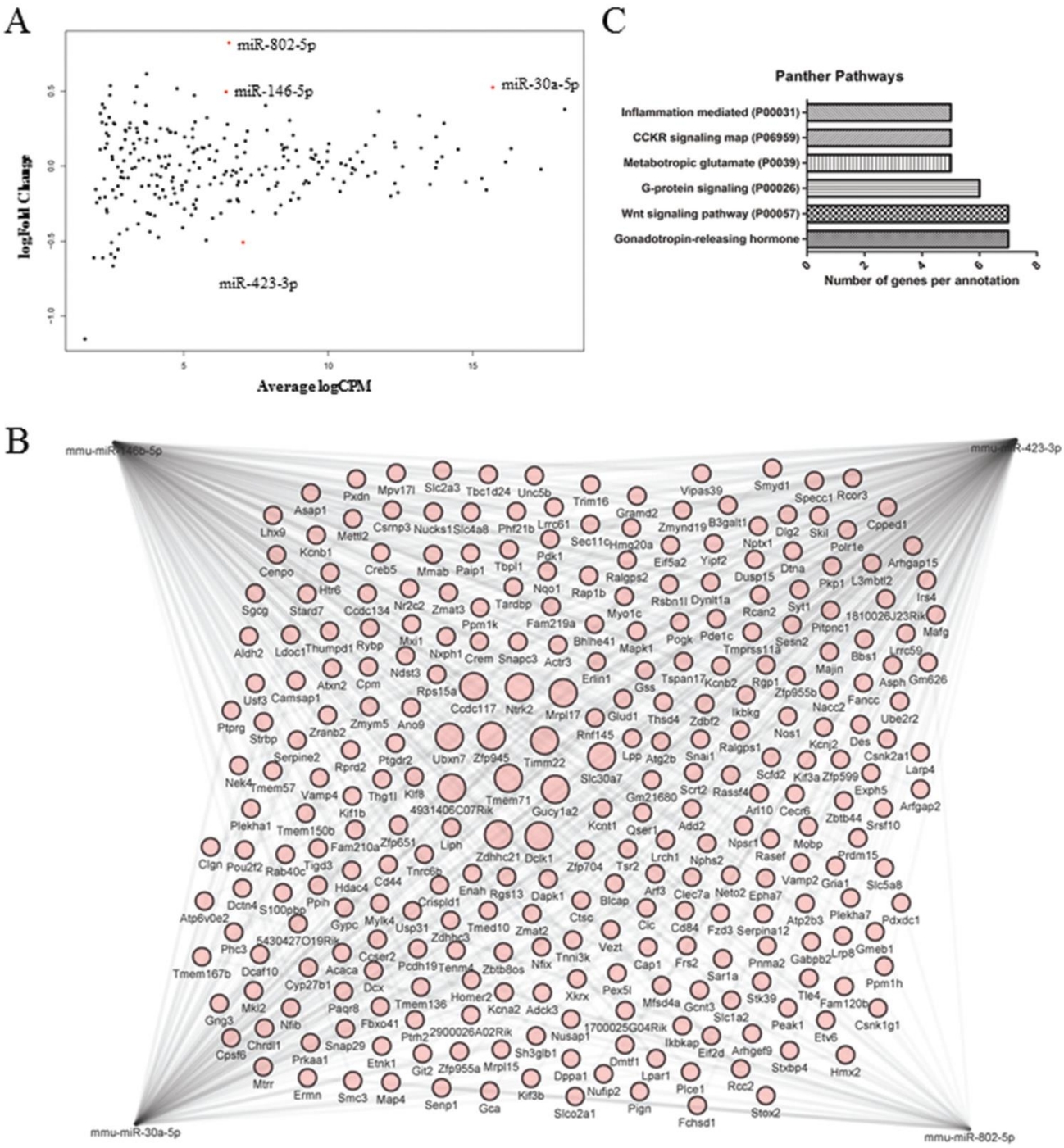


Fig. 5 Liver miRNA analysis. (A) Scatter plot of the RNA-seq data of liver miRNAs from mice supplemented with HT, for 8 weeks. (B) Genetic interaction analysis between miRNAs and their likely miRNA targets. Target point sizes are directly correlated with the number of interactions within the set of miRNAs. (C) Functional enrichment analysis of differentially expressed miRNA targets. HT, hydroxytyrosol.

establish functional connections among compounds triggering similar responses at the molecular level by using computational approaches involving machine learning tools such as hierarchical clustering.[41] For example, we can predict the pharmacological properties of many molecules across different biological systems and conditions solely based on their transcriptional profiles.[41] Yet, there are few applications of such approaches in the emerging field of nutrigenomics, which investigates the effects of food and nutrients on gene expression.[12] Here, we analyzed the available transcriptomic and proteomic data to (1) identify differentially expressed genes and proteins prevailing among studies addressing

hydroxytyrosol supplementation *in vivo* and (2) validate the identified differentially expressed genes and proteins as robust targets of HT.

We first searched for *in vivo* experiments involving dietary supplementation with HT where high throughput gene and protein expression data were generated. Then, we identified a signature of dozens of genes shared among the selected studies which could be related to the biological effects associated with HT consumption. Although several genes exhibited different expressions in at least two tissues, only four transcripts were significantly modulated in the four tissues we analyzed, *i.e.* brain, adipose tissue, liver, and intestine. This finding seemed to be particularly relevant, considering that the data were obtained from two independent laboratories[25,26] and microarrays platforms. Regarding differentially expressed proteins, very few candidates (less than five) were identified as commonly modulated in at least two different tissues (heart, aortic, hepatic, and adipose tissues) and none was common to all tissues. Proteomic data were generated from three different laboratories.[22,42,43] *In silico* analysis allowed us to identify 18 genes and a reduced number of proteins, which were subsequently tested for validation in rodent liver samples. Knowing that the liver is crucial for lipid metabolism and that samples were available from all studies, our hypothesis was that consistent modulation of specific targets of HT could be found in this tissue in animals supplemented with HT.

The transcriptomics studies included in this work were performed in five different tissues: breast, adipose tissue, intestine, liver, and brain. Among the four commonly differentially regulated genes selected for validation, *Efr3a* was down-regulated in all tissues, except for the intestine. *Plscr1* was downregulated in all tissues, whereas *Kctd2* was only downregulated in the intestine and liver and *Slc37a4* was upregulated in the brain and adipose tissue. Validation of these four genes in the liver samples from the selected studies (Studies 1–4) confirmed the upregulation of *Slc37a4* in Studies 2 and 3. The other genes did not change significantly, suggesting a large inter-study variability. None of the 14 additionally selected genes for validation (Fig. 1) changed in all four studies. While some genes only changed in one study (*Anks6*, *Plscr2*, *Lipe*, *Snx16*, *Ppp1cb*, *Top1* and *Tip2*), others changed in two studies (*Sort1*, *Slc37a4*, *Rora* and *Fgf21*). According to our results, Study 4 showed the most reduced changes in gene expression, but we do not know whether this is related to the fact that this study was performed in rats rather than in mice. Moreover, in contrast with the other three mice studies receiving HT, in this study rats received secoiridoids.[22] Secoiridoids are the major precursors of HT, after their *in vivo* digestion.[2] Regarding the function of these genes, *Sort1* influences plasma lipid concentration.[44] *Slc37a4* is a glucose-6-phosphate translocase (G6PT), which transports G6P from the cytoplasm into the endoplasmic reticulum (ER) lumen, and is involved in glucose metabolism.[45] *Rora* is a nuclear receptor involved in multiple biological processes, including lipid metabolism.[46] *Fgf21* is a metabolic gene that influences plasma glucose and triglyceride levels,[47] and is a critical regulator of liver lipid homeostasis.[48] Moreover, FGF21 is induced directly by PPARalpha in the liver in response to fasting[49] and its induction is required for the normal activation of hepatic lipid oxidation and triglyceride clearance.[48] As such, induced expression of *Fgf21* by HT feeding could be beneficial against metabolic diseases. Because *Fgf21* and *Rora* are important contributors to metabolic diseases, we further validated their response to HT supplementation in a different cohort.[20] Interestingly, C57Bl6J mice receiving a single ingestion of HT dramatically increased their hepatic expression of *Fgf21* at 1, 2, and 4 h post-ingestion. This effect was also observed for *Rora*, but to a lower degree 4 h post-ingestion (Fig. 6). In this sense, our bioinformatic approach and further validation uncovered novel possible molecular targets of the beneficial effects of HT consumption. A previous study performed in a different mouse model showed that the repression of *Fgf21* caused by a high fat diet was reverted by HT supplementation.[50] These and our current data suggest *Fgf21* to be a bona fide candidate target of HT. Whether this effect occurs in humans is unknown and deserves further investigation.

In the study by Tomé-Carneiro *et al.*, 2017 where samples deriving from Study 1 were used, a decrease in the expression of FASN and PRDX1 was recorded, by both high throughput proteomics and WB, in mice supplemented with HT for eight weeks (Tomé-Carneiro *et al.*, 2017).[42] However, WB revealed only non-significant changes for these proteins in the liver samples tested for validation. WB analysis also revealed non-significant changes for CAR3 in the studies used for validation, despite it being reported as upregulated by high throughput proteomics in Study 1.

According to high throughput proteomics, VIM was significantly changed in the adipose tissue samples from Study 1 and the aortic tissue from Study 4. Here, however, WB analysis

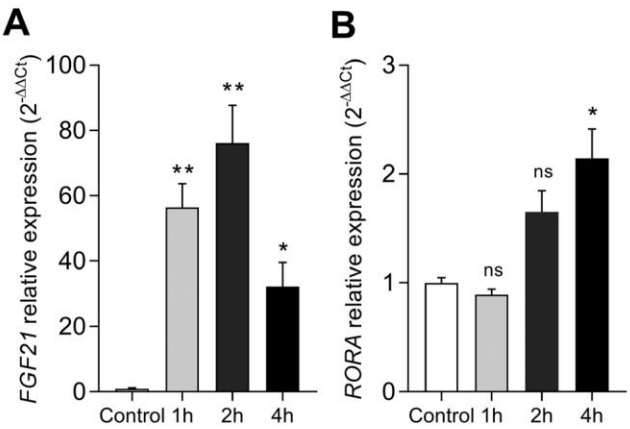


Fig. 6 Hydroxytyrosol target liver mRNA expression of *Fgf21* and *Rora*. Effects of hydroxytyrosol on *FGF21* (A) and RORA and (B) relative gene expression in liver samples of C57Bl6J mice at different time points (1, 2 and 4 hours). Data shown as mean ± SEM. * $p < 0.05$ compared to the control group; ** $p < 0.0001$ compared to the control group. ($n = 9$ per group).

showed non-significant changes for these proteins in the liver samples used for validation. The statistically significant changes observed for ACTN4 proteins in liver (Study 1), aortic, and heart tissues (Study 4), by WB, were not confirmed in the liver samples used for validation. Likewise, the statistically significant changes observed for HSPD1 in adipose (Study 1), hepatic (Study 3) and aortic (Study 4) tissues were not confirmed, by WB, in the liver samples used for validation. Overall, none of the proteins selected for validation showed consistent differential expression across the liver samples tested here.

It is important to note that the transcriptional data analysis was performed independently of the protein data levels and that correlations between transcripts and proteins were not intended during the validation process. Also, it is relevant to mention that the list of common genes or proteins to be validated were common to at least two different studies, regardless of whether the tissues subjected to high throughput analysis matched or not. Indeed, in most cases, tissues from where transcript (intestine, liver, adipose, brain or breast tumor) and protein (liver, heart, aortic, or adipose) levels were selected did not match. Thus, the lack of consistent validation of potential targets of HT in response to dietary supplementation should be seen with caution.

Other aspects of the complex regulatory variation from RNA to protein may account for the lack of common tissue features in response to HT supplementation. For instance, studies in model organisms and humans have shown that variations in mRNA and protein expression levels are often uncorrelated.[51,52] Few transcripts are exclusive to a particular tissue and varies more across tissues than individuals,[53] while genetic variation can also influence the heterogenicity of the protein expression in a diverse set of human tissues.[52,54] Moreover, differences may also arise from alterations in post-translational regulation. Regulation by ncRNAs, particularly miRNAs, is among the plethora of posttranslational controlling pathways.

Although in the past few years increasing evidence has suggested that food bioactive compounds can modulate the expression of miRNAs *in vitro*,[55] in animal models,[56] and in humans,[57] very few studies have specifically focused on the action of HT. For example, specific miRNAs, miR-9[58] and miR-146a,[59] were evaluated *in vitro*, whereas only one study evaluated the whole miRNome in the mouse small intestine.[20] Among the liver modulated miRNAs in response to HT supplementation, miR-802-5p has been previously described as being obesity-inducing and as being involved in glucose metabolism impairment and in angiotensin signaling regulation.[60,61] As for miR-423-3p, its levels have been positively associated with cell growth in liver, colon or other types of cancers.[62,63] Induction of miR-30a-5p has been previously described to ameliorate liver fibrosis[64] or to suppress breast tumor growth and metastasis.[64,65] miR-146b has been shown to attenuate non-alcoholic steatohepatitis,[66] although its down-regulation has been shown to promote cancer growth and metastasis.[67]

Sustained intake of HT at dietary doses by mice resulted in altered miRNA expression in the intestine (assessed in Study 1) and the liver (assessed here). Of note, HT supplementation resulted only in a consistent ($p < 0.05$) regulation of miR-802-5p in both tissues (ESI Fig. S1†). The reduced number of common modulated miRNAs found could be explained by different aspects of miRNA biogenesis, function, and technical analysis. For example, some miRNAs are tissue-specific.[68] Also, though the processing pattern of miRNAs in tissues and cell lines may differ, it has been reported that, especially in cell lines, several transcribed miRNAs are not processed to mature miRNA.[69] While miRNAs might play a role in the biological action of HT, the very reduced number of HT studies evaluating miRNAs precludes any conclusion regarding their regulatory potential at this point. However, the consistent induction of miR-802-5p in two different tissues, in response to dietary HT supplementation, seems to support a miRNA modulating action of small natural molecules, which could be exploited as a potential therapeutic alternative or adjuvant to the current pharmacological arsenal targeting endogenous miRNAs.

5 Conclusions

High throughput transcriptomics and proteomics are powerful tools that greatly contribute to the knowledge of how nutrition affects the expression of a wide number of genes and proteins. Although *in vivo* studies where these techniques are employed to investigate the molecular effects of hydroxytyrosol are increasing in number, they are still scarce. Therefore, we believe that there is a growing need to integrate the accumulating data in order to identify the consistent targets of this bioactive compound. Most of the genes and proteins identified and tested here as potential HT targets showed inconsistent modulation by HT. These results are, at least in part, due to the limited number of *in vivo* studies available, with heterogeneous HT doses and supplementation times, where different tissues were used for transcriptomic/proteomic analysis. However, our transcriptomic analysis uncovered two novel potential HT target candidates, *i.e. Fgf21* and *Rora*. While we certainly do not want to depreciate the important role of omics and their related database, we feel that more attention should be paid to the current pitfalls of this approach to nutritional research. Over-emphasis should be avoided and more HT-supplementation studies employing high throughput transcriptomics and proteomics tools are needed for potential HT targets to be identified and validated.

Author contributions

MCLH, RMH and MCC contributed equally to this work. AD, MCLH and FV contributed to the conception or design of the work. RMH, MCC, JT-C, MB R-R and LdP contributed to data collection. LR, MJM, JO and MN contributed with samples from different studies. AD, FV, JAM and MPP obtained finan-

cial support. MCC, JT-C, MCLH, RMH, AD and FV drafted the article. JAM, JO, JCE-G and MPP revised the manuscript for important intellectual contribution. All authors reviewed and approved the manuscript.

Abbreviations

HT Hydroxytyrosol
GO Gene ontology
GI Gene Interaction

Conflicts of interest

The authors declare no conflicts of interest related to this work.

Acknowledgements

This research was funded by grants from the Spanish "Agencia Estatal de Investigación" and the European FEDER Funds to AD and RMH (AGL2016-78922-R); the Ministerio de Economía y Competitividad–Fondo Europeo de Desarrollo Regional (SAF2016-75441-R) and the Fondo Social Europeo–Gobierno de Aragón (B16_17R) to MAN and JO; the Fundación Ramón Areces (CIVP18A3888) to AD, JG-Z, JTC, MCC, FV and RMH; the CIBER de Fisiopatología de la Obesidad y Nutrición (CIBERObn) to JAM, MPP, JO and MAN; and POR FESR 3S4H to FV. CIBERObn. is an initiative of the ISCIII, Spain. MCLH, LdP and MB R-R were recipients of contracts from the Consejería de Educación, Juventud y Deporte de la Comunidad de Madrid, Fondo Social Europeo, and the Iniciativa de Empleo Juvenil YEI (PEJD-2016/BIO-2781, PEJD-2017-PRE/BIO-5100, and PEJD-2018-POST/BIO 8933), respectively.

References

1 M. C. L. de las Hazas, L. Rubio, A. Macia and M. J. Motilva, Hydroxytyrosol: Emerging Trends in Potential Therapeutic Applications, *Curr. Pharm. Des.*, 2018, **24**, 2157–2179.
2 M.-C. López de las Hazas, C. Piñol, A. Macià, M.-P. Romero, *et al.*, Differential absorption and metabolism of hydroxytyrosol and its precursors oleuropein and secoiridoids, *J. Funct. Foods*, 2016, **22**, 52–63.
3 L. Rubió, R.-M. Valls, A. Macià, A. Pedret, *et al.*, Impact of olive oil phenolic concentration on human plasmatic phenolic metabolites, *Food Chem.*, 2012, **135**, 2922–2929.
4 M. de Bock, E. B. Thorstensen, J. G. B. Derraik, H. V. Henderson, *et al.*, Human absorption and metabolism of oleuropein and hydroxytyrosol ingested as olive (Olea europaea L.) leaf extract, *Mol. Nutr. Food Res.*, 2013, **57**, 2079–2085.
5 O. Khymenets, M. C. Crespo, O. Dangles, N. Rakotomanomana, *et al.*, Human hydroxytyrosol's absorption and excretion from a nutraceutical, *J. Funct. Foods*, 2016, **23**, 278–282.
6 F. Visioli, A. Davalos, M. López de las Hazas, M. C. Crespo, *et al.*, An overview of the pharmacology of olive oil and its active ingredients, *Br. J. Pharmacol.*, 2019, DOI: 10.1111/bph.14782.
7 M. Crespo, J. Tomé-Carneiro, A. Dávalos and F. Visioli, Pharma-Nutritional Properties of Olive Oil Phenols. Transfer of New Findings to Human Nutrition, *Foods*, 2018, **7**, 90.
8 R. Valenzuela, F. Echeverria, M. Ortiz, M.Á Rincón-Cervera, *et al.*, Hydroxytyrosol prevents reduction in liver activity of Δ-5 and Δ-6 desaturases, oxidative stress, and depletion in long chain polyunsaturated fatty acid content in different tissues of high-fat diet fed mice, *Lipids Health Dis.*, 2017, **16**, 64.
9 S. Tejada, S. Pinya, M. Mar Bibiloni, J. A. del Tur, *et al.*, Cardioprotective Effects of the Polyphenol Hydroxytyrosol from Olive Oil, *Curr. Drug Targets*, 2017, **18**, 1477–1486.
10 N. Richard, S. Arnold, U. Hoeller, C. Kilpert, *et al.*, Hydroxytyrosol is the major anti-inflammatory compound in aqueous olive extracts and impairs cytokine and chemokine production in macrophages, *Planta Med.*, 2011, **77**, 1890–1897.
11 A. Pedret, S. Fernández-Castillejo, R.-M. Valls, Ú. Catalán, *et al.*, Cardiovascular Benefits of Phenol-Enriched Virgin Olive Oils: New Insights from the Virgin Olive Oil and Hdl Functionality (Vohf) Study, *Mol. Nutr. Food Res.*, 2018, 1800456.
12 R. Martín-Hernández, G. Reglero and A. Dávalos, Data mining of nutrigenomics experiments: Identification of a cancer protective gene signature, *J. Funct. Foods*, 2018, **42**, 380–386.
13 P. Suravajhala, L. J. A. Kogelman and H. N. Kadarmideen, Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare, *Genet., Sel., Evol.*, 2016, **48**, 38.
14 D. Braconi, G. Bernardini, L. Millucci and A. Santucci, Foodomics for human health: current status and perspectives, *Expert Rev. Proteomics*, 2018, **15**, 153–164.
15 C. Pavlidis, Z. Lanara, A. Balasopoulou, J.-C. Nebel, *et al.*, Meta-Analysis of Genes in Commercially Available Nutrigenomic Tests Denotes Lack of Association with Dietary Intake and Nutrient-Related Pathologies, *OMICS: J. Integr. Biol.*, 2015, **19**, 512–520.
16 P. A. Maranhão, G. M Bacelar-Silva, D. N. G. Ferreira, C. Calhau, *et al.*, Nutrigenomic Information in the openEHR Data Set, *Appl. Clin. Inform.*, 2018, **9**, 221–231.
17 F. Vitali, R. Lombardo, D. Rivero, F. Mattivi, *et al.*, ONS: an ontology for a standardized description of interventions and observational studies in nutrition, *Genes Nutr.*, 2018, **13**, 12.
18 S. Dhanasekaran, T. K. Bhattacharya, R. N. Chatterjee, C. Paswan, *et al.*, Functional genomics in chicken (Gallus gallus) - status and implications in poultry, *Worlds. Poult. Sci. J.*, 2014, **70**, 45–56.

19 R. Agarwal and V. Dhar, Editorial-Big Data, Data Science, and, Analytics: The Opportunity and Challenge for IS Research, *Inf. Syst. Res.*, 2014, **25**, 443–448.
20 J. Tomé-Carneiro, M. C. Crespo, E. Iglesias-Gutierrez, R. Martín, *et al.*, Hydroxytyrosol supplementation modulates the expression of miRNAs in rodents and in humans, *J. Nutr. Biochem.*, 2016, **34**, 146–155.
21 S. Acín, M. A. Navarro, J. M. Arbonés-Mainar, N. Guillén, *et al.*, Hydroxytyrosol administration enhances atherosclerotic lesion development in apo E deficient mice, *J. Biochem.*, 2006, **140**, 383–391.
22 Ú. Catalán, L. Rubió, M.-C. López de las Hazas, P. Herrero, *et al.*, Hydroxytyrosol and its complex forms (secoiridoids) modulate aorta and heart proteome in healthy rats: Potential cardio-protective effects, *Mol. Nutr. Food Res.*, 2016, **60**, 2114–2129.
23 C. Sticht, C. De La Torre, A. Parveen and N. Gretz, miRWalk: An online resource for prediction of microRNA binding sites, *PLoS One*, 2018, **13**, e0206239.
24 H. Mi, X. Huang, A. Muruganujan, H. Tang, *et al.*, PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements, *Nucleic Acids Res.*, 2017, **45**, D183–D189.
25 S. Granados-Principal, J. L. Quiles, C. Ramirez-Tortosa, P. Camacho-Corencia, *et al.*, Hydroxytyrosol inhibits growth and cell proliferation and promotes high expression of sfrp4 in rat mammary tumours, *Mol. Nutr. Food Res.*, 2011, **55**(Suppl 1), S117–S126.
26 E. Giordano, A. Dávalos and F. Visioli, Chronic hydroxytyrosol feeding modulates glutathione-mediated oxido-reduction pathways in adipose tissue: A nutrigenomic study, *Nutr., Metab. Cardiovasc. Dis.*, 2014, **24**, 1144–1150.
27 E. C. Lai, Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation, *Nat. Genet.*, 2002, **30**, 363–364.
28 D. Baek, J. Villén, C. Shin, F. D. Camargo, *et al.*, The impact of microRNAs on protein output, *Nature*, 2008, **455**, 64–71.
29 J. M. Ordovas and D. Corella, Nutritional genomics., *Annu. Rev. Genomics Hum. Genet.*, 2004, **5**, 71–118.
30 A. P. Simopoulos, Nutrigenetics/Nutrigenomics, *Annu. Rev. Public Health*, 2010, **31**, 53–68.
31 H. Nakayama, Y. Shimada, L. Zang, M. Terasawa, *et al.*, Novel Anti-Obesity Properties of Palmaria mollis in Zebrafish and Mouse Models, *Nutrients*, 2018, **10**, 1401.
32 A. Alkhatib, C. Tsang and J. Tuomilehto, Olive Oil Nutraceuticals in the Prevention and Management of Diabetes: From Molecules to Lifestyle, *Int. J. Mol. Sci.*, 2018, **19**, 2024.
33 J. Logan and M. W. Bourassa, The rationale for a role for diet and nutrition in the prevention and treatment of cancer, *Eur. J. Cancer Prev.*, 2018, **27**, 406–410.
34 M.-H. Pan, J.-C. Wu, C.-T. Ho and C.-S. Lai, Antiobesity molecular mechanisms of action: Resveratrol and pterostilbene, *BioFactors*, 2018, **44**, 50–60.
35 L. Väremo, C. Scheele, C. Broholm, A. Mardinoglu, *et al.*, Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes, *Cell Rep.*, 2015, **11**, 921–933.
36 V. do Nascimento de Oliveira, A. B. M. Lima-Neto, M. F. van Tilburg, A. C. de Oliveira Monteiro-Moreira, *et al.*, Proteomic analysis to identify candidate biomarkers associated with type 1 diabetes, *Diabetes, Metab. Syndr. Obes.*, 2018, **11**, 289–301.
37 X. Yan, Y. Wu, F. Zhong, Q. Jiang, *et al.*, iTRAQ and PRM-based quantitative proteomics in T2DM-susceptible and -tolerant models of Bama mini-pig, *Gene*, 2018, **675**, 119–127.
38 M. Murri, M. Insenser, M. R. Bernal-Lopez, P. Perez-Martinez, *et al.*, Proteomic analysis of visceral adipose tissue in pre-obese patients with type 2 diabetes, *Mol. Cell. Endocrinol.*, 2013, **376**, 99–106.
39 B. de Roos and D. F. Romagnolo, Proteomic Approaches to Predict Bioavailability of Fatty Acids and Their Influence on Cancer and Chronic Disease Prevention, *J. Nutr.*, 2012, **142**, 1370S–1376S.
40 Y. E. Kim, H. J. Jeon, D. Kim, S. Y. Lee, *et al.*, Quantitative Proteomic Analysis of 2D and 3D Cultured Colorectal Cancer Cells: Profiling of Tankyrase Inhibitor XAV939-Induced Proteome, *Sci. Rep.*, 2018, **8**, 13255.
41 A. Aliper, S. Plis, A. Artemov, A. Ulloa, *et al.*, Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data, *Mol. Pharm.*, 2016, **13**, 2524–2530.
42 J. Tomé-Carneiro, M. C. Crespo, E. García-Calvo, J. L. Luque-García, *et al.*, Proteomic evaluation of mouse adipose tissue and liver following hydroxytyrosol supplementation, *Food Chem. Toxicol.*, 2017, **107**, 329–338.
43 G. Rodríguez-Gutiérrez, G. G. Duthie, S. Wood, P. Morrice, *et al.*, Alperujo extract, hydroxytyrosol, and 3,4-dihydroxyphenylglycol are bioavailable and have antioxidant properties in vitamin E-deficient rats–a proteomics and network analysis approach, *Mol. Nutr. Food Res.*, 2012, **56**, 1137–1147.
44 C. J. Willer, S. Sanna, A. U. Jackson, A. Scuteri, *et al.*, Newly identified loci that influence lipid concentrations and risk of coronary artery disease, *Nat. Genet.*, 2008, **40**, 161–169.
45 A. R. Cappello, R. Curcio, R. Lappano, M. Maggiolini, *et al.*, The Physiopathological Role of the Exchangers Belonging to the SLC37 Family, *Front. Chem.*, 2018, **6**, 122.
46 K. Kim, K. Boo, Y. S. Yu, S. K. Oh, *et al.*, RORα controls hepatic lipid homeostasis via negative regulation of PPARγ transcriptional network, *Nat. Commun.*, 2017, **8**, 162.
47 A. Kharitonenkov, T. L. Shiyanova, A. Koester, A. M. Ford, *et al.*, FGF-21 as a novel metabolic regulator, *J. Clin. Invest.*, 2005, **115**, 1627–1635.
48 M. K. Badman, P. Pissios, A. R. Kennedy, G. Koukos, *et al.*, Hepatic Fibroblast Growth Factor 21 Is Regulated by PPARα and Is a Key Mediator of Hepatic Lipid Metabolism in Ketotic States, *Cell Metab.*, 2007, **5**, 426–437.
49 T. Inagaki, P. Dutchak, G. Zhao, X. Ding, *et al.*, Endocrine Regulation of the Fasting Response by PPARα-Mediated Induction of Fibroblast Growth Factor 21, *Cell Metab.*, 2007, **5**, 415–425.

50 C. Pirozzi, A. Lama, R. Simeoli, O. Paciello, *et al.*, Hydroxytyrosol prevents metabolic impairment reducing hepatic inflammation and restoring duodenal integrity in a rat model of NAFLD, *J. Nutr. Biochem.*, 2016, **30**, 108–115.

51 E. J. Foss, D. Radulovic, S. A. Shaffer, D. M. Ruderfer, *et al.*, Genetic basis of proteome variation in yeast, *Nat. Genet.*, 2007, **39**, 1369–1375.

52 A. Battle, Z. Khan, S. H. Wang, A. Mitrano, *et al.*, Impact of regulatory variation from RNA to protein, *Science*, 2015, **347**, 664–667.

53 M. Mele, P. G. Ferreira, F. Reverter, D. S. DeLuca, *et al.*, The human transcriptome across tissues and individuals, *Science*, 2015, **348**, 660–665.

54 GTEx Consortium, Gte., Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, *Science*, 2015, **348**, 648–660.

55 L. Baselga-Escudero, C. Blade, A. Ribas-Latre, E. Casanova, *et al.*, Resveratrol and EGCG bind directly and distinctively to miR-33a and miR-122 and modulate divergently their levels in hepatic cells, *Nucleic Acids Res.*, 2014, **42**, 882–892.

56 D. Milenkovic, C. Deval, E. Gouranton, J.-F. Landrier, *et al.*, Modulation of miRNA Expression by Dietary Polyphenols in apoE Deficient Mice: A New Mechanism of the Action of Polyphenols, *PLoS One*, 2012, **7**, e29837.

57 M. A. Nuñez-Sánchez, A. Dávalos, A. González-Sarrías, P. Casas-Agustench, *et al.*, MicroRNAs expression in normal and malignant colon tissues as biomarkers of colorectal cancer and in response to pomegranate extracts consumption: Critical issues to discern between modulatory effects and potential artefacts, *Mol. Nutr. Food Res.*, 2015, **59**, 1973–1986.

58 S. D'Adamo, S. Cetrullo, S. Guidotti, R. M. Borzì, *et al.*, Hydroxytyrosol modulates the levels of microRNA-9 and its target sirtuin-1 thereby counteracting oxidative stress-induced chondrocyte death, *Osteoarthritis Cartilage*, 2017, **25**, 600–610.

59 E. Bigagli, L. Cinci, S. Paccosi, A. Parenti, *et al.*, Nutritionally relevant concentrations of resveratrol and hydroxytyrosol mitigate oxidative burst of human granulocytes and monocytes and the production of pro-inflammatory mediators in LPS-stimulated RAW 264.7 macrophages, *Int. Immunopharmacol.*, 2017, **43**, 147–155.

60 S. E. Sansom, G. J. Nuovo, M. M. Martin, S. R. Kotha, *et al.*, miR-802 regulates human angiotensin II type 1 receptor expression in intestinal epithelial C2BBe1 cells, *Am. J. Physiol.: Gastrointest. Liver Physiol.*, 2010, **299**, G632–G642.

61 J.-W. Kornfeld, C. Baitzel, A. C. Könner, H. T. Nicholls, *et al.*, Obesity-induced overexpression of miR-802 impairs glucose metabolism through silencing of Hnf1b, *Nature*, 2013, **494**, 111–115.

62 H.-T. Li, H. Zhang, Y. Chen, X.-F. Liu, *et al.*, MiR-423–3p enhances cell growth through inhibition of p21Cip1/Waf1 in colorectal cancer, *Cell. Physiol. Biochem.*, 2015, **37**, 1044–1054.

63 J. Lin, S. Huang, S. Wu, J. Ding, *et al.*, MicroRNA-423 promotes cell growth and regulates G 1 /S transition by targeting p21Cip1/Waf1 in hepatocellular carcinoma, *Carcinogenesis*, 2011, **32**, 1641–1647.

64 J. Chen, Y. Yu, S. Li, Y. Liu, *et al.*, MicroRNA-30a ameliorates hepatic fibrosis by inhibiting Beclin1-mediated autophagy, *J. Cell. Mol. Med.*, 2017, **21**, 3679–3692.

65 L. Li, L. Kang, W. Zhao, Y. Feng, *et al.*, miR-30a-5p suppresses breast tumor growth and metastasis through inhibition of LDHA-mediated Warburg effect, *Cancer Lett.*, 2017, **400**, 89–98.

66 W. Jiang, J. Liu, Y. Dai, N. Zhou, *et al.*, MiR-146b attenuates high-fat diet-induced non-alcoholic steatohepatitis in mice, *J. Gastroenterol. Hepatol.*, 2015, **30**, 933–943.

67 C. Li, R. Miao, S. Liu, Y. Wan, *et al.*, Down-regulation of miR-146b-5p by long noncoding RNA MALAT1 in hepatocellular carcinoma promotes cancer growth and metastasis, *Oncotarget*, 2017, **8**, 28683–28695.

68 M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, *et al.*, Identification of tissue-specific microRNAs from mouse, *Curr. Biol.*, 2002, **12**, 735–739.

69 E. J. Lee, M. Baek, Y. Gusev, D. J. Brackett, *et al.*, Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors, *RNA*, 2008, **14**, 35–42.

70 A. Voigt, J. Ribot, A. G. Sabater, A. Palou, *et al.*, Identification of Mest/Peg1 gene expression as a predictive biomarker of adipose tissue expansion sensitive to dietary anti-obesity interventions, *Genes Nutr.*, 2015, **10**, 27.

71 A. Zheng, H. Li, J. Xu, K. Cao, *et al.*, Hydroxytyrosol improves mitochondrial function and reduces oxidative stress in the brain of db/db mice: role of AMP-activated protein kinase activation, *Br. J. Nutr.*, 2015, **113**, 1667–1676.

72 A. Zheng, H. Li, K. Cao, J. Xu, *et al.*, Maternal hydroxytyrosol administration improves neurogenesis and cognitive function in prenatally stressed offspring, *J. Nutr. Biochem.*, 2015, **26**, 190–199.

DATABASE
The Journal of Biological Databases and Curation

Database, 2019, 1–9
doi: 10.1093/database/baz097
Original article

Original article

NutriGenomeDB: a nutrigenomics exploratory and analytical platform

AQ1 **Roberto Martín-Hernández[1,*], Guillermo Reglero[2,3], José M Ordovás[4,5] and Alberto Dávalos[6]**

AQ2 [1] Bioinformatics and Biostatistics Unit, IMDEA Food Institute, CEI UAM+CSIC, Ctra. De Canto Blanco 8, Madrid 28049, Spain, [2]Sección Departamental de Ciencias de la Alimentación, Facultad de Ciencias, Universidad Autónoma de Madrid, Ctra. De Canto Blanco 8, Madrid 28049, Spain, [3]Laboratory of Food Products for Precision Nutrition, IMDEA Food Institute, CEI UAM+CSIC, Ctra. De Canto Blanco 8,
AQ3 Madrid 28049, Spain, [4] Nutrition and Genomics Laboratory, JM-USDA Human Nutrition Research Center on Aging, Tufts University, Boston, MA 02111, USA, [5]Laboratory of Nutritional Genomics, IMDEA Food Institute, CEI UAM+CSIC, Ctra. De Canto Blanco 8, Madrid 280149, Spain and [6]Laboratory of Epigenetics of Lipid Metabolism, IMDEA Food Institute, CEI UAM+CSIC, Ctra. De Canto Blanco 8, Madrid 28049, Spain

AQ4 *Corresponding author: Tel: +34917278100; Email: roberto.martin@imdea.org

Citation details: Martín-Hernández,R., Reglero,G., Ordovás,J.M. *et al.* NutriGenomeDB: a nutrigenomics exploratory and analytical platform. *Database* (2019) Vol. 2019: article ID baz097; doi:10.1093/database/baz097

Received 12 April 2019; Revised 3 June 2019; Accepted 1 July 2019

Abstract

Habitual consumption of certain foods has shown beneficial and protective effects against multiple chronic diseases. However, it is not clear by which molecular mechanisms they may exert their beneficial effects. Multiple -omic experiments available in public databases have generated gene expression data following the treatment of human cells with different food nutrients and bioactive compounds. Exploration of such data in an integrative manner offers excellent possibilities for gaining insights into the molecular effects of food compounds and bioactive molecules at the cellular level. Here we present NutriGenomeDB, a web-based application that hosts manually curated gene sets defined from gene expression signatures, after differential expression analysis of nutrigenomics experiments performed on human cells available in the Gene Expression Omnibus (GEO) repository. Through its web interface, users can explore gene expression data with interactive visualizations. In addition, external gene signatures can be connected with nutrigenomics gene sets using a gene pattern-matching algorithm. We further demonstrate how the application can capture the primary molecular mechanisms of a drug used to treat hypertension and thus connect its mode of action with hosted food compounds.
Database URL: http://nutrigenomedb.org.

Introduction

Nutrigenomics is defined as the science studying the role of nutrients and bioactive food compounds on gene ex pression. Research in nutrition and food technologies is attracting interest from the scientific community, as reflected by the increasing number of published scientific works related to this field since the last 10 years. There is evidence pointing out that nutrition may exert its impact on health outcomes by directly affecting the expression of genes in critical metabolic pathways. Research from the '80s already suggested that diet likely accounted for about 30% of the risk of developing cancer (1). Moreover, diet is considered as one of the main risk factors in the etiology of cardiovascular diseases (CVDs). It is widely accepted that dietary components can regulate cellular processes spanning-from gene expression to protein synthesis, but there is a minimal understanding of the underlying molecular mechanisms (2).

The endless increasing number of transcriptomic data generated with high-throughput technologies, which are available in public repositories, has triggered the development of web applications that allow the comparison of gene expression profiles for distinct purposes. The Connectivity Map (CMap) (3) was the first introduction of such a methodology, aiming at connecting small molecules, genes and diseases. Nowadays, their catalogue of cellular signatures representing perturbations with pharmacologic perturbagens keeps growing. NFFinder is another similar application that makes use of transcriptomic data for drug repurposing issues in the context of orphan diseases (4). CREEDS is also an interesting project based on the same principle but focusing on crowdsourcing manual curation of gene signatures instead of one-fits-all automated gene expression analysis (5).

Despite the increasing availability of nutrigenomics experiments in public databases, such data remain underutilized, and therefore the information about modes of action of food compounds and natural products that might explain their observed healthy properties at the cellular level is scarce (6). For instance, the existence of software packages tailored to analyse and mine data generated from nutritional research is almost inexistent, and researchers working in this field are restricted to analyse their experiments individually. Previous computational analysis of food-derived bioactive compounds has allowed the identification of possible biological function or molecular mechanism. For example, molecular docking simulation has been used to evaluate in silico inhibitors of cyclooxygenase 2 (7) and pancreatic lipase (8), or BET bromodomains (9) by natural dietary bioactive compounds. In the context of a whole diet, computational frameworks using gene expression profiles have been used to establish diet-disease associations (10), which could be used for designing future dietary recommendations. Natural products present a high structural heterogeneity and are a rich source of drugs or drug-like lead (11). Although the final products may not necessarily represent the active ingredients of the natural source, a large number of drugs in the market have their origin in nature. Since the effect of a drug treatment on a cell culture can be captured by the triggered gene expression profile, a resource allowing to establish connections among different food compounds, as well as connections between food compounds and drugs available in the market, based on the similarity of their effect on gene expression, would be highly valuable.

Here we present NutriGenomeDB, an easy-to-use web application that allows exploration of differential gene expression profiles from nutrigenomics experiments through data tables and interactive visualizations. The phenotype-centred analysis module allows the comparison of human expression profiles against a curated built-in database of nutrigenomics experiments performed in human cells. The application is freely accessible at http://nutrigenomedb.org.

Materials and Methods

Data collection

Gene expression profiles were identified in the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo). Queries were conducted with variations of the following search terms: nutrigenomics, nutrient, nutrition, extract, natural product and phytochemical. In general, selected studies were performed on human cellular models, using microarray technology, and the experimental design included at least two replicates per group.

Data processing

Samples were manually assigned to control or treatment condition by inspecting the corresponding experimental design of each study. For data corresponding to Affymetrix platforms, raw data (CEL files) were downloaded and normalized locally with the Robust Multi-array Average algorithm using specific Bioconductor packages. For data generated with other microarray platforms, the normalized matrix was directly downloaded from the source for differential expression analysis. Differential gene expression was assessed using the LIMMA package from Bioconductor R project and setting the treated samples as the target group. Each experiment was characterized by a gene set defined as the top 10% differentially expressed genes, avoiding any filtering steps based on statistical significance. These gene

OUP UNCORRECTED PROOF – FIRST PROOF, 30/7/2019, SPi

sets were then sorted by the level of differential expression (log2 fold change). Conversion of probes to gene names was performed locally with the 'annotate' Bioconductor library.

Web interface

NutriGenomeDB was built using Ruby on Rails framework. Gene differential expression data for each experiment is hosted in a MySQL database. The database can be searched with one or multiple gene symbols in order to get gene expression results among the obtained nutrigenomics gene sets. Interactive tables presenting the results were implemented using DataTables JavaScript and are available for download in either Excel or PDF format.

Two different open source libraries were used to implement interactive visualizations on the gene-centred exploratory module. After interrogating the stored gene sets for a particular set of genes, the available multiline plot visualization, allowing exploration of gene expression levels across nutrigenomics experiments, was implemented using Plotly's Python graphing library. The expression heatmap functionality, able to cluster nutrigenomics experiments based on the differential expression profile of a set of query genes, was built using the Clustergrammer web-based tool using specific JavaScript and Python libraries (12). For gene signature comparison functionality, the Gene Set Enrichment Analysis algorithm was implemented (13), using the obtained nutrigenomics gene sets as a reference database.

AQ6

Submitted queries on the phenotype-centred analysis module are identified with a unique job ID and remain stored for later accession on the web server for 1 month. Identified genes connecting the experiments can be used to launch a statistical overrepresentation test of molecular functions via PANTHER database web services.

Results

NutriGenomeDB overview

NutriGenomeDB contains manually curated gene sets defining nutrigenomics experiments available in the GEO database. Based on the selection criteria, a total of 61 nutrigenomics studies were identified in the GEO database. The experimental design of these studies represents 231 differential gene expression experiments (Supplementary Table 1). Therefore, each experiment is defined by a gene set, representing around the top 10% differentially expressed genes sorted by the level of differential expression. Gene sets range from 1500 to 3000 gene features depending on the platform used for gene expression analysis. From a total of 568.463 stored rows in NutriGenomeDB, 156.374 are statistically significant at an adjusted *P*-value equal or lower than 0.05, corresponding to 27.51% of statistically significant genes among the hosted nutrigenomics gene sets. The gathered data was generated among 19 distinct microarray platforms. NutriGenomeDB is based on two main modules: (i) a data browse module allowing exploration of gene expression data across experiments through interactive visualizations and (ii) a phenotype-centred analysis module allowing the comparison of gene signatures, aimed at finding potential connections among phenotypes and food compounds (Figure 1). The essence of the phenotype-centred analysis module approach is to quantify the similarity of external gene expression signatures characterizing a phenotype in response to a specific treatment (gene expression results from users) and those triggered by food bioactive components available in the literature (GEO), by quantifying such a connection through a gene signature pattern-matching algorithm. Identified connections might be quantified either by the total number of overlapping genes or the calculated net enrichment score (NES). This score is obtained by walking down the introduced list of genes, ranked by their level of differential expression, increasing a running-sum statistic when a gene belongs to the nutrigenomics gene set and decreasing it when the gene does not (13). Such an increase in the running sum is proportional to the rank metric for that gene, so the contribution of highly over- and underexpressed genes will be more important. The enrichment score is the maximum deviation from zero encountered during that walk. NES is the corresponding adjusted enrichment score, which accounts for differences in the gene set sizes included in NutriGenomeDB. A highly positive NES means that overlapping genes between the query and reference gene signatures are mainly found on the top of both lists (highly overexpressed in both cases), and vice versa for a negative NES.

Exploring nutrigenomics gene expression

The exploratory module requires at least one human gene symbol as input. NutriGenomeDB outputs an interactive data table which includes gene differential expression information, such as the calculated log2 fold change together with statistical data, and summary information about the source experiment, such as the tested food compound and additional experimental design data. Each experiment is linked to the source data through their GEO's ID. The data table is sortable and can be filtered by searching specific food compounds/nutrients, GEO's ID or any other experimental data. The presented results only show the experiments where the queried genes appear in the gene set defining the analyzed nutrigenomics experiments. An interactive line plot can be generated using the queried genes and can be downloaded as a PNG image (Figure 2).

OUP UNCORRECTED PROOF – FIRST PROOF, 30/7/2019, SPi

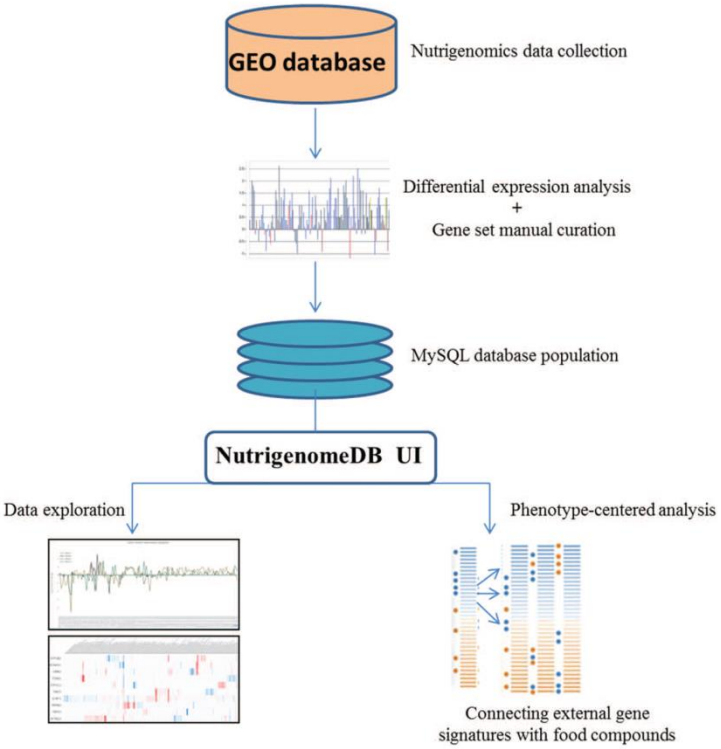


Figure 1. Framework of the methodology for setting up NutriGenomeDB web application. Following identification of studies related to nutrigenomics at the cellular level, experiments were analyzed for gene differential expression. Manually curated gene sets defining each experiment with specific foods and bioactive compound were defined and stored in a MySQL database. The web application interface allows performing exploratory analysis of those data through interactive visualizations. An analysis module allows connecting external gene signatures with food compounds using a gene pattern-matching algorithm.

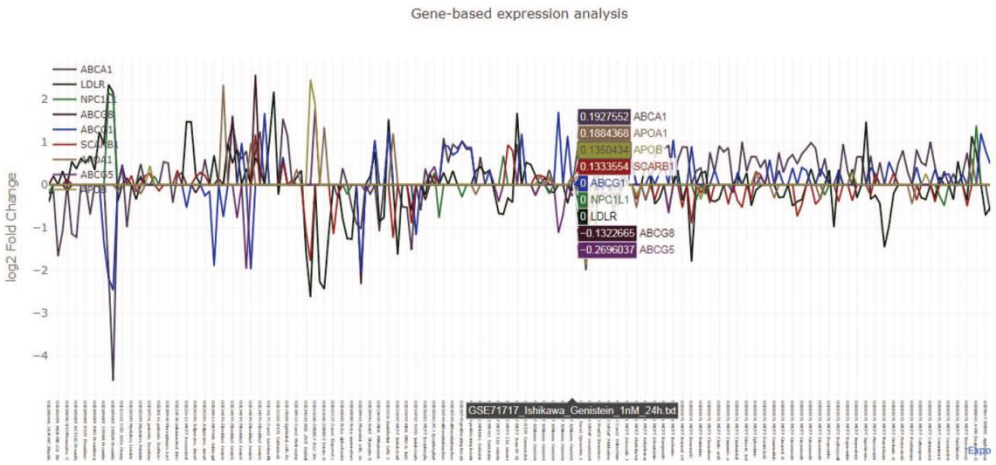
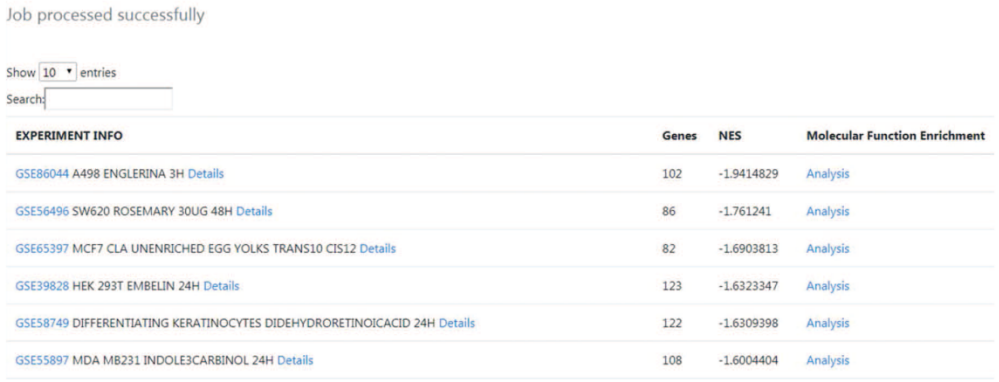


Figure 2. Graphical output of the nutrigenomics gene expression exploratory module. The query is composed of 12 key genes involved in cholesterol metabolism (ABCA1, ABCG1, ABCG5, ABCG8, NPC1L1, APOB, APOA1, LDLR, NPC1L1, APOA1, NR1H1, SCARB1). The obtained line plot is interactive and allows visual identification of co-expression patterns. In this example, treatment of Ishikawa cells with 100 nM of genistein for 24 h triggers the upregulation of LDLR and SCARB1 and the downregulation of ABCA1 and ABCG5.

OUP UNCORRECTED PROOF – FIRST PROOF, 30/7/2019, SPi


Job processed successfully

Show 10 entries

Search:

EXPERIMENT INFO	Genes	NES	Molecular Function Enrichment
GSE86044 A498 ENGLERINA 3H Details	102	-1.9414829	Analysis
GSE56496 SW620 ROSEMARY 30UG 48H Details	86	-1.761241	Analysis
GSE65397 MCF7 CLA UNENRICHED EGG YOLKS TRANS10 CIS12 Details	82	-1.6903813	Analysis
GSE39828 HEK 293T EMBELIN 24H Details	123	-1.6323347	Analysis
GSE58749 DIFFERENTIATING KERATINOCYTES DIDEHYDRORETINOICACID 24H Details	122	-1.6309398	Analysis
GSE55897 MDA MB231 INDOLE3CARBINOL 24H Details	108	-1.6004404	Analysis

Figure 3. Output results from the phenotype-centred analysis module. A list of nutrigenomics experiments connected with the introduced gene signature is obtained. It included information about the number of overlapping genes (Genes), the NES and experimental information directly linked to the source data. For each connection, details about the expression level and statistics of the overlapping genes between gene signatures can be inspected by clicking on the 'Details' link. A molecular function enrichment analysis of the overlapping genes can be launched by clicking on the 'Analysis' link.

As nutrigenomics gene sets were defined by the top 10% differentially expressed genes sorted by their level of differential expression, genes that are not included in all the nutrigenomics gene sets show a log2 fold change value of 0 in those where they are absent.

Connecting external gene signatures with food compounds

This module is aimed at identifying connections between nutrigenomics experiments and introduced gene expression signatures characterizing specific phenotypes. The idea behind this approach is that food compounds and nutrients may exert their potential health benefits by acting on the same molecular targets that drugs used for treating diseases do. The introduced data must contain gene symbols and quantitative log2 fold change expression information. It is compared against the gene sets included in NutriGenomeDB by gene pattern-matching. Results are presented as a data table informing about matched nutrigenomics experiments, the number of overlapping genes between the introduced query and experiments hosted in the database, and information about whether those overlapping genes are mostly up- or downregulated [normalized enrichment score (NES) column]. In order to get insights into the molecular mechanisms underlying the identified connections, a molecular function enrichment analysis of those genes connecting both experiments can be launched from the user interface.

To illustrate the utility of this analysis module, we decided to use a gene signature triggered by amlodipine, a drug used to treat high blood pressure. An experiment testing the effect of amlodipine on the gene expression of human umbilical vein endothelial cells (HUVECs) was found and downloaded for analysis (GEO ID: GSE42808). From the list of differentially expressed genes sorted by statistical significance, the first 1000 features were used to query the NutriGenomeDB analysis tool.

The results table highlights connections between the query gene signature and nutrigenomics experiments performed with compounds such as englerin A, a rosemary extract, and *trans*-10,*cis*-12 conjugated linoleic acid. The table was sorted by the strongest negative connection level (NES score), in order to obtain connections with compounds mostly leading to downregulation of genes similar to the amlodipine drug profile (Figure 3).

To get further insights into the molecular mechanisms explaining these results, a molecular function enrichment analysis was performed from the results page ('Analysis' button on the right of each row). Thus, the overlapping genes between the introduced gene expression profile and the nutrigenomics gene sets were used as input data. Even

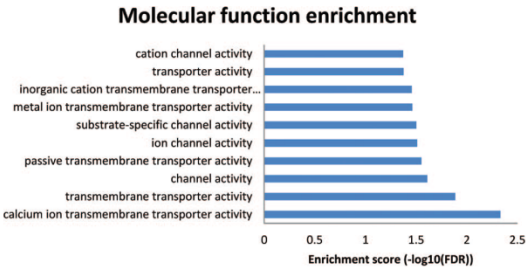


Figure 4. Results from the enriched molecular functions obtained from the NutriGenomeDB phenotype-centred analysis module. The presented bar chart includes the overrepresented molecular functions with statistical significance (FDR < 0.05).

Table 1. Genes commonly downregulated between the query gene signature and connected nutrigenomics experiment (GSE56496)

Gene symbol	log2FC amlodipine	log2FC GSE56496 rosemary 30 µg 48 h	Description
TRPV2	−1.327098208	−0.674827345	Transient receptor potential cation channel subfamily V member 2
ATP2C2	−1.413693464	−1.119488698	ATPase secretory pathway Ca^{2+} transporting 2
GRIK2	−1.329016735	−1.027980433	Glutamate ionotropic receptor kainate type subunit 2
SORCS3	−2.566320697	−3.12378339	Sortilin-related VPS10 domain containing receptor 3

though the first experiment on the results list (GSE86044 englerin A treatment for 3 h on A498 cells) did not show any overrepresented molecular function with statistical significance, the results from the 86 overlapping genes of the second experiment (GSE56496) performed with a rosemary extract were promising (Figure 4). The enriched molecular functions showing statistical significance are tightly related to transmembrane transporter activities of calcium ions. Indeed, these results are in good agreement with the molecular mechanism of amlodipine drug, which acts as a calcium channel blocker (https://pubchem.ncbi.nlm.nih.gov/compound/Amlodipine).

This analysis module also allows users to inspect the gene expression profile of genes connecting the experiments by clicking on the 'Details' link. As expected, among the genes related to transmembrane-transported activities as molecular functions, some of them were downregulated together in the amlodipine gene signature and the Rosemary extract gene set (Table 1).

With the aim of finding in NutriGenomeDB additional food compounds able to modulate the identified genes related to transmembrane transporter activities, the heatmap tool from the gene exploratory module was used. We used as input a list of 11 genes identified with such a molecular function. Thus, we were able to identify a cluster comprising 16 different experiments which modulated at least one of those genes. Interestingly, the ATPase secretory pathway Ca^{2+} transporting 2 gene (ATP2B2) was consistently downregulated in 11 out of the 16 clustered experiments, with an important downregulation in the experiment GSE34074, corresponding to a treatment using 100 µM of genistein in fibroblasts for 24 h (Figure 5). Interestingly, several animal model studies support genistein's potential as an anti-hypertensive agent (14–16). However, clinical trials for genistein are very limited, and those for soy isoflavones do not support their contribution to the lowering of blood pressure in humans (17, 18).

Discussion

Since the first introduction of the CMap principle and methodology (3), there have been numerous applications of this approach by many research groups. However, the focus has been mostly limited to drug discovery and

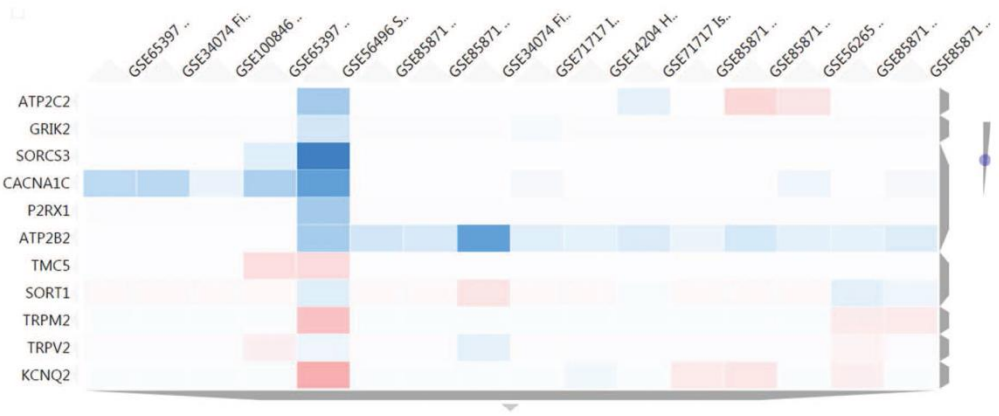


Figure 5. Heatmap results using as input a set of 11 genes related to transmembrane transporter activities. The figure shows a cluster of 16 nutrigenomics experiments based on the expression level of those genes (blue colour represents underexpression, vice versa for red colour). ATP2B2 shows a consistent downregulation among 11 of the clustered experiments.

OUP UNCORRECTED PROOF – FIRST PROOF, 30/7/2019, SPi

repurposing issues. Also, existing tools that automatically compute gene signatures from the GEO database in a high-throughput manner (19, 20) have demonstrated to be error-prone because of the automatic assignment of control and perturbation samples, together with automatic signature generation and annotation. Thus, a database with manually curated gene signatures extracted from nutrigenomics experiments appears to be reliable and highly valuable for the research community.

We have demonstrated, through a benchmark analysis of the gene signature caused by the amlodipine drug on human cells, that our application can capture the main molecular mechanisms responsible for the drug mode of action and to further link these data with food compounds able to trigger those molecular mechanisms. By comparing both gene signatures, we found that an experiment performed using a Rosemary extract obtained from the GEO database (GSE56496) is closely related to the molecular effect of amlodipine. Indeed, there are numerous references in the scientific literature about the benefits of rosemary for the treatment of hypertension (21–23). Interestingly, following a heatmap analysis of the connecting genes that are involved in transmembrane transporter activities, we found a cluster comprising food compounds such as genistein, which has demonstrated strong hypotensive effects (24). Therefore, the newly defined and manually curated nutrigenomics gene sets are able to capture similarities in the compound mode of action rather than similarities of the experimental settings. Thus, we believe that the NutriGenomeDB tool can be helpful to generate a new research hypothesis with the aim of contributing to the field of precision nutrition, improving future functional food formulations, and further investigate the molecular mechanisms that confer healthy properties to specific dietary components.

The biological effects of nutrients and food bioactives derived from diet depend on physiological processes such as absorption, transportation, binding to the cell's nuclear receptors and excretion. Therefore, the comprehensive identification of bioactive compounds from food components that are responsible for the observed health benefits, through analytical techniques, is a critical step in order to improve the bioavailability of molecules and thus to benefit from their healthy properties. NutriGenomeDB hosts experiments that might not seem obvious from a nutritional point of view, such as the study GSE74212, which investigates the effect of the ascidian natural product eusynstyelamide B, a novel topoisomerase II poison that induces DNA damage and growth arrest in prostate and breast cancer cells. However, at appropriate concentrations, such a natural product could have potential applications for functional foods with protective effects. Indeed, sea squirts are eaten by humans in many parts of the world, including Japan, Korea, Chile and Europe.

The aim of NutriGenomeDB is to become a central hub of nutrigenomics data and to remain on constant development, continuously adding newly generated datasets and features to increase the connectivity power.

Previous studies on drug repositioning using computational approaches have revolutionized the discovery of new uses for existing drugs (25). This has led to the development of several computational tools (26, 27), some of which uses transcriptomic data (28) or even adverse reaction database (29), among others. However, the use of computational approaches for food bioactive, nutritional or nutrigenomics studies is scarce. Perhaps one of the aspects that prevent comparative analysis of food bioactive transcriptomic effects is the lack of standardization of nutrigenomics data in a meaningful way to be reusable. This highlights the need for using minimum information about a bioactive entity (30) as those now used by the pharmaceutical industry. As diet is a major player for maintaining health and preventing disease, the use of system biology approaches not only will open up new opportunities to decipher the health-promoting properties of foods but also may provide new opportunities towards personalization of nutrition (31). Indeed, in a recent study using publicly available gene expression profiles for foods, disease and drugs, Zheng and colleagues provided a diet–disease association using system-level interactomics analysis and network-based inference to provide dietary recommendations.

The lack of similar nutritional systems biology approach in the context of food bioactive compounds prompted us to develop this manually curated database resource. Previous studies by our group described the analysis of a large amount of transcriptomic data and its curation (6), which is the basis for this web application. To the best of our knowledge, this is the first of its kind in the field of nutritional genomics, but because of the relevance of system biology in nutritional sciences (10), we envision the development of other applications.

Concluding Remarks

Diet is a major player of health and disease, and their cellular response at transcriptomic levels may provide important clues for understanding their molecular mode of action. The increasing number of large-scale public gene expression data (transcriptomic) and experiments on dietary food bioactive compounds provide a unique opportunity to interrogate molecules with a similar biological response at the gene expression level. In this context, NutriGenomeDB, a manually curated web-based application of nutrigenomics experiments, provides a unique tool to the nutritional field for identifying mech-

anistic similarities in the mode of action of food bioactive compounds and other bioactive molecules. The essence of NutriGenomeDB is to quantify the similarity of gene expression signatures between an external gene signature (users query) and those of food bioactive components available in the literature (GEO). This database could be useful for both the nutrition and pharmaceutical industry to search for molecules that share the same molecular effects at the gene level. NutriGenomeDB can be easily extended to include various organisms and incorporate novel GEO transcriptomic data. Understanding the biological effects at gene expression levels of food bioactive components not only provides a platform for searching bioactive molecules with similar biological effects but also may provide novel predictive tools for the development of functional foods and personalization of diet.

Author Contributions

R.M.-H. and A.D. designed the study. R.M.-H. performed the study and analyzed the data. R.M.-H. and A.D. wrote the main manuscript text. All authors reviewed the manuscript.

Supplementary Data

Supplementary data are available at *Database* Online.

AQ7

Funding

Fundación Ramón Areces (CIVP18A3888 to A.D. and R.M.-H.); Spanish Agencia Estatal de Investigación and the European Fonds Européen de Développement Économique et Régional Funds (AGL2016-78922-R to A.D. and R.M.-H.); US Department of Agriculture (under agreement no. 8050-51000-098-00D to J.M.O).

Conflict of interest. None declared.

References

1. Doll,R. and Peto,R. (1981) The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J. Natl. Cancer. Inst.*, **66**, 1191–1308.
2. Panagiotou,G. and Nielsen,J. (2009) Nutritional systems biology: definitions and approaches. *Annu. Rev. Nutr.*, **29**, 329–339.
3. Lamb,J., Crawford,E.D., Peck,D. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
4. Setoain,J., Franch,M., Martinez,M. *et al.* (2015) NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. *Nucleic Acids Res.*, **43**, W193–W199.
5. Wang,Z., Monteiro,C.D., Jagodnik,K.M. *et al.* (2016) Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.*, **7**, 12846.
6. Martín-Hernández,R., Reglero,G. and Dávalos,A. (2018) Data mining of nutrigenomics experiments: identification of a cancer protective gene signature. *J. Funct. Foods*, **42**, 380–386.
7. Maldonado-Rojas,W. and Olivero-Verbel,J. (2011) Potential interaction of natural dietary bioactive compounds with COX-2. *J. Mol. Graph Model.*, **30**, 157–166.
8. Birari,R.B., Gupta,S., Mohan,C.G. *et al.* (2011) Antiobesity and lipid lowering effects of Glycyrrhiza chalcones: experimental and computational studies. *Phytomedicine*, **18**, 795–801.
9. Dutra,L.A., Heidenreich,D., Silva,G. *et al.* (2017) Dietary compound resveratrol is a pan-BET bromodomain inhibitor. *Nutrients*, **9**. AQ8
10. Zheng,T., Ni,Y., Li,J. *et al.* (2017) Designing dietary recommendations using system level Interactomics analysis and network-based inference. *Front Physiol.*, **8**, 753.
11. Batova,A., Altomare,D., Creek,K.E. *et al.* (2017) Englerin A induces an acute inflammatory response and reveals lipid metabolism and ER stress as targetable vulnerabilities in renal cell carcinoma. *PloS One*, **12**, e0172632.
12. Fernandez,N.F., Gundersen,G.W., Rahman,A. *et al.* (2017) Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci. Data*, **4**, 170151.
13. Subramanian,A., Tamayo,P., Mootha,V.K. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.
14. Zheng,Z., Yu,S., Zhang,W. *et al.* (2017) Genistein attenuates monocrotaline-induced pulmonary arterial hypertension in rats by activating PI3K/Akt/eNOS signaling. *Histol. Histopathol.*, **32**, 35–41.
15. Sun,L., Zhao,T., Ju,T. *et al.* (2015) A combination of intravenous genistein plus Mg2+ enhances antihypertensive effects in SHR by endothelial protection and BKCa channel inhibition. *Am. J. Hypertens.*, **28**, 1114–1120.
16. Matori,H., Umar,S., Nadadur,R.D. *et al.* (2012) Genistein, a soy phytoestrogen, reverses severe pulmonary hypertension and prevents right heart failure in rats. *Hypertension*, **60**, 425–430.
17. Teede,H.J., Giannopoulos,D., Dalais,F.S. *et al.* (2006) Randomised, controlled, cross-over trial of soy protein with isoflavones on blood pressure and arterial function in hypertensive subjects. *J. Am. Coll. Nutr.*, **25**, 533–540.
18. Hodgson,J.M., Puddey,I.B., Beilin,L.J. *et al.* (1999) Effects of isoflavonoids on blood pressure in subjects with high-normal ambulatory blood pressure levels: a randomized controlled trial. *Am. J. Hypertens.*, **12**, 47–53.
19. Zinman,G.E., Naiman,S., Kanfi,Y. *et al.* (2013) ExpressionBlast: mining large, unstructured expression databases. *Nat. Methods*, **10**, 925–926.
20. Williams,G. (2013) SPIEDw: a searchable platform-independent expression database web tool. *BMC Genomics*, **14**, 765.
21. Hassani,F.V., Shirani,K. and Hosseinzadeh,H. (2016) Rosemary (Rosmarinus officinalis) as a potential therapeutic plant in metabolic syndrome: a review. *Naunyn Schmiedebergs Arch. Pharmacol.*, **389**, 931–949.
22. Apostolidis,E., Kwon,Y.I. and Shetty,K. (2006) Potential of cranberry-based herbal synergies for diabetes and hypertension management. *Asia Pac. J. Clin. Nutr.*, **15**, 433–441.
23. Neves,J.A. and Oliveira,R.C.M. (2018) Pharmacological and biotechnological advances with Rosmarinus officinalis L. *Expert Opin. Ther. Pat.*, **28**, 399–413.

24. Sureda,A., Sanches Silva,A., Sanchez-Machado,D.I. *et al.* (2017) Hypotensive effects of genistein: from chemistry to medicine. *Chem. Biol. Interact.*, **268**, 37–46.
25. Chaudhari,R., Tan,Z., Huang,B. *et al.* (2017) Computational polypharmacology: a new paradigm for drug discovery. *Expert Opin. Drug Discov.*, **12**, 279–291.
26. Napolitano,F., Carrella,D., Mandriani,B. *et al.* (2018) gene2drug: a computational tool for pathway-based rational drug repositioning. *Bioinformatics*, **34**, 1498–1505.
27. Wan,F., Hong,L., Xiao,A. *et al.* (2019) NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics*, **35**, 104–111.
28. Wang,Y., Yella,J. and Jegga,A.G. (2019) Transcriptomic data mining and repurposing for computational drug discovery. *Methods Mol. Biol.*, **1903**, 73–95.
29. Oh,M., Ahn,J., Lee,T. *et al.* (2017) Drug voyager: a computational platform for exploring unintended drug action. *BMC Bioinformatics*, **18**, 131.
30. Orchard,S., Al-Lazikani,B., Bryant,S. *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.*, **10**, 661–669.
31. Badimon,L., Vilahur,G. and Padro,T. (2017) Systems biology approaches to understand the effects of nutrition and promote health. *Br. J. Clin. Pharmacol.*, **83**, 38–45.